2021

# Applications Of Random Matrix Theory In Statistics And Machine Learning

Yue Sheng
*University of Pennsylvania*

# Applications Of Random Matrix Theory In Statistics And Machine Learning


## Abstract

We live in an age of big data. Analyzing modern data sets can be very difficult because they usually present the following features: massive, high-dimensional, and heterogeneous. How to deal with these new features often plays a key role in modern statistical and machine learning research. This dissertation uses random matrix theory (RMT), a powerful mathematical tool, to study several important problems where the data is massive, high-dimensional, and sometimes heterogeneous.

The first chapter briefly introduces some basics of random matrix theory (RMT). We also cover some classical applications of RMT to statistics and machine learning.

The second chapter is about distributed linear regression, where we consider the ordinary least squares (OLS) estimators. Distributed statistical learning problems arise commonly when dealing with large datasets. In this setup, datasets are partitioned over machines, which compute locally and communicate short messages. Communication is often the bottleneck. We study one-step and iterative weighted parameter averaging in statistical linear models under data parallelism. We do linear regression on each machine, send the results to a central server, and take a weighted average of the parameters. Optionally, we iterate, sending back the weighted average and doing local ridge regressions centered at it. How does this work compare to doing linear regression on the full data? Here we study the performance loss in estimation and test error, and confidence interval length in high dimensions, where the number of parameters is comparable to the training data size. We find the performance loss in one-step weighted averaging, and also give results for iterative averaging. We also find that different problems are affected differently by the distributed framework.

The third chapter studies a fundamental and highly important problem in this area: How to do ridge regression in a distributed computing environment? Ridge regression is an extremely popular method for supervised learning and has several optimality properties, thus it is important to study. We study one-shot methods that construct weighted combinations of ridge regression estimators computed on each machine. By analyzing the mean squared error in a high dimensional random-effects model where each predictor has a small effect, we discover several new phenomena. We also propose a new Weighted ONe-shot DistributEd Ridge regression (WONDER) algorithm. We test WONDER in simulation studies and using the Million Song Dataset as an example. There it can save at least 100x in computation time, while nearly preserving test accuracy.

The fourth chapter is trying to solve another possible issue with modern data sets, that is heterogeneity. Dimensionality reduction via PCA and factor analysis is an important tool of data analysis. A critical step is selecting the number of components. However, existing methods (such as the scree plot, likelihood ratio, parallel analysis, etc) do not have statistical guarantees in the increasingly common setting where the data are heterogeneous. There each noise entry can have a different distribution. To address this problem, we propose the Signflip Parallel Analysis (Signflip PA) method: it compares data singular values to those of “empirical null” data generated by flipping the sign of each entry randomly with probability one-half. We show that Signflip PA consistently selects factors above the noise level in high-dimensional signal-plus-noise models (including spiked models and factor models) under heterogeneous settings. Here the classical parallel analysis is no longer effective. To do this, we propose to leverage recent breakthroughs in random matrix theory, such as dimension-free operator norm bounds and large deviations for the top eigenvalues of nonhomogeneous matrices. We also illustrate that Signflip PA performs well in numerical simulations and on empirical data examples.

# APPLICATIONS OF RANDOM MATRIX THEORY IN STATISTICS AND MACHINE LEARNING

Yue Sheng

A DISSERTATION

in

Applied Mathematics and Computational Science

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2021

Supervisor of Dissertation

---

Edgar Dobriban, Assistant Professor of Statistics

Co-Supervisor of Dissertation

---

Robin Pemantle, Merriam Term Professor of Mathematics

Graduate Group Chairperson

---

Robin Pemantle, Merriam Term Professor of Mathematics

Dissertation Committee
Edward George, Universal Furniture Professor in Statistics and Economics

APPLICATIONS OF RANDOM MATRIX THEORY IN STATISTICS AND MACHINE LEARNING

# ACKNOWLEDGMENT

First and foremost, I would like to thank my main advisor Professor Edgar Dobriban. I appreciate his contributions of time and ideas for my research. Without his generous support, the completion of this dissertation would have not been possible. I would like to thank Professor Robin Pemantle for co-advising me. Although I did not have the opportunity to work with you, I am truly grateful for all the warm conversations and kind suggestions.

I would like to thank Professor Edward George for serving on my dissertation committee. I would also like to express my gratitude to my wonderful collaborators David Hong and Fan Yang for the stimulating discussions and inspiring comments. I also cannot imagine to have any accomplishment without the academic and non-academic discussions with many friends, including Mauricio Daros Andrade, Jorge Barreras, Zhiqi Bu, Kan Chen, Shuxiao Chen, Zongyu Dai, Jinshuo Dong, Sheng Gao, Yansong Gao, Siyu Heng, Xinyu Liao, Mingyang Liu, Lingxi Lu, Sagnik Nandy, Hongming Pu, Hua Wang, Xuran Wang, Yichen Wang, Jing Xu, Yachong Yang, Bo Zhang, Yiliang Zhang, Linjun Zhang, and many others. My sincere thanks also go to Professor Charles Epstein and Professor Pedro Ponte Castañeda, the former graduate chairs of AMCS, and the program coordinators, for their continuous support.

I am deeply grateful to my wife and my parents for their dedicated love and support ever since I met them. Finally, I would like to thank my own spirit for not giving up during the hardest days and making the dream come true.

# ABSTRACT

## APPLICATIONS OF RANDOM MATRIX THEORY IN STATISTICS AND MACHINE LEARNING

Yue Sheng

Edgar Dobriban

Robin Pemantle

We live in an age of big data. Analyzing modern data sets can be very difficult because they usually present the following features: massive, high-dimensional, and heterogeneous. How to deal with these new features often plays a key role in modern statistical and machine learning research. This dissertation uses random matrix theory (RMT), a powerful mathematical tool, to study several important problems where the data is massive, high-dimensional, and sometimes heterogeneous.

The first chapter briefly introduces some basics of random matrix theory (RMT). We also cover some classical applications of RMT to statistics and machine learning.

The second chapter is about distributed linear regression, where we consider the ordinary least squares (OLS) estimators. Distributed statistical learning problems arise commonly when dealing with large datasets. In this setup, datasets are partitioned over machines, which compute locally, and communicate short messages. Communication is often the bottleneck. We study one-step and iterative weighted parameter averaging in statistical linear models under data parallelism. We do linear regression on each machine, send the results to a central server, and take a weighted average of the parameters. Optionally, we iterate, sending back the weighted average and doing local ridge regressions centered at it. How does this work compared to doing linear regression on the full data? Here we study the performance loss in estimation and test error, and confidence interval length in high dimensions, where the number of parameters is comparable to the training data size. We find the performance loss in one-step weighted averaging, and also give results for iterative

averaging. We also find that different problems are affected differently by the distributed framework.

The third chapter studies a fundamental and highly important problem in this area: How to do ridge regression in a distributed computing environment? Ridge regression is an extremely popular method for supervised learning, and has several optimality properties, thus it is important to study. We study one-shot methods that construct weighted combinations of ridge regression estimators computed on each machine. By analyzing the mean squared error in a high dimensional random-effects model where each predictor has a small effect, we discover several new phenomena. We also propose a new Weighted ONe-shot DistributEd Ridge regression (WONDER) algorithm. We test WONDER in simulation studies and using the Million Song Dataset as an example. There it can save at least 100x in computation time, while nearly preserving test accuracy.

The fourth chapter is trying to solve another possible issue with modern data sets, that is heterogeneity. Dimensionality reduction via PCA and factor analysis is an important tool of data analysis. A critical step is selecting the number of components. However, existing methods (such as the scree plot, likelihood ratio, parallel analysis, etc) do not have statistical guarantees in the increasingly common setting where the data are heterogeneous. There each noise entry can have a different distribution. To address this problem, we propose the Signflip Parallel Analysis (Signflip PA) method: it compares data singular values to those of "empirical null" data generated by flipping the sign of each entry randomly with probability one-half. We show that Signflip PA consistently selects factors above the noise level in high-dimensional signal-plus-noise models (including spiked models and factor models) under heterogeneous settings. Here classical parallel analysis is no longer effective. To do this, we propose to leverage recent breakthroughs in random matrix theory, such as dimension-free operator norm bounds and large deviations for the top eigenvalues of nonhomogeneous matrices. We also illustrate that Signflip PA performs well in numerical simulations and on empirical data examples.

TABLE OF CONTENTS

ACKNOWLEDGMENT . . . . . iii

ABSTRACT . . . . . iv

LIST OF TABLES . . . . . viii

LIST OF ILLUSTRATIONS . . . . . xii

CHAPTER 1 : A Brief Introduction to Random Matrix Theory and Its Applications 1

CHAPTER 2 : Distributed Linear Regression by Averaging . . . . . 5

2.1 Introduction . . . . . 5

2.2 Some related work . . . . . 8

2.3 One-step weighted averaging: General linear functionals . . . . . 11

2.4 Calculus of deterministic equivalents . . . . . 16

2.5 Examples . . . . . 21

2.6 Insights for Parameter Estimation . . . . . 31

2.7 Multi-shot methods . . . . . 31

2.8 Appendix . . . . . 41

CHAPTER 3 : WONDER: Weighted One-shot Distributed Ridge Regression in High Dimensions . . . . . 75

3.1 Introduction . . . . . 75

3.2 Finite Sample Results . . . . . 82

3.3 Asymptotics under Linear Random-effects Models . . . . . 87

3.4 Special Case: Identity Covariance . . . . . 94

3.5 WONDER: Algorithms for Weighted One-shot Distributed Ridge Regression 109

3.6 Experimental Results . . . . . 111

3.7 Appendix . . . . . 116

CHAPTER 4 : Selecting the Number of Components in PCA via Random Signflips 143

4.1 Introduction . . . . . 143

4.2 Parallel analysis, heterogeneous noise, and the need for new methods . . . . 148

4.3 Proposed method: Signflip parallel analysis . . . . . 154

4.4 Theoretical analysis and guarantees . . . . . 156

4.5 Implications for rank selection . . . . . 177

4.6 Experiments . . . . . 181

4.7 Results on signals . . . . . 190

4.8 Proofs regarding the noise . . . . . 200

4.9 Proofs regarding rank selection . . . . . 211

BIBLIOGRAPHY . . . . . 212

# LIST OF TABLES

TABLE 1 : Estimation, Confidence Interval, and test efficiency as a function of number of machines $k$, the sample size $n$, and the dimension $p$. This is how much smaller the error of the global estimator is compared to the distributed estimator. These functions are plotted and described in Figure 1. . . . . 8

TABLE 2 : A general framework for finite-sample efficiency calculations. The rows show the various statistical problems studied in our work, namely estimation, confidence interval formation, in-sample prediction, out-of-sample prediction and regression function estimation. The elements of the row show how these tasks fall in the framework of linear functional prediction described in the main body. . . . . 13

# LIST OF ILLUSTRATIONS

FIGURE 1 : How much accuracy do we lose in distributed regression? The plots show the relative efficiency, i.e., the ratio of errors, of the global least squares (OLS) estimator, compared to the distributed estimator averaging the local least squares estimators. This efficiency is at most unity, because the global estimator is more accurate. If the efficiency is close to unity, then averaging is accurate. We show the behavior of estimation and test error, as a function of number of machines. We see that estimation error is *much more affected* than test error. The specific formulas are given in Table 1. . . . . 5

FIGURE 2 : Plots of $\eta$ and $f$ for $G$ being the point mass at unity. . . . . 22

FIGURE 3 : Comparison of empirical and theoretical ARE for standard sample covariance matrices. Left: $n = 10,000$, $p = 20$. Right: $n = 10,000$, $p = 100$. . . . . 25

FIGURE 4 : Relative efficiency for Marchenko-Pastur model. . . . . 28

FIGURE 5 : Comparison of the one-shot weighted method and several widely used multi-shot methods, . . . . . 41

FIGURE 6 : Comparison of empirical and theoretical ARE for elliptical distributions. Left: $n = 10,000$, $p = 20$. Right: $n = 10,000$, $p = 100$. . . . . 55

FIGURE 7 : NYC flights data. . . . . 73

FIGURE 8 : Efficiency loss due to one-shot distributed learning. This plot shows the relative mean squared error of centralized ridge regression compared to optimally weighted one-shot distributed ridge regression. This quantity is at most unity, and the larger, the "better" distributed ridge works. Specifically, the model is asymptotic, and we show the dependence of the Asymptotic Relative Efficiency (ARE) on the aspect ratio $\gamma = \lim p/n$ (where $n$ is sample size and $p$ is dimension) and on the signal strengh $\alpha = \sqrt{\mathbb{E}\|\beta\|^2}$, in the *infinite-worker limit* when we distribute our data over many machines. We show (a) surface and (b) contour plots of the ARE. See the text for details. 78

FIGURE 9 : Plots of the optimal risk function $\phi$ as a function of the aspect ratio $\gamma$ (denoted by $x$ in the plots), for different signal strength parameters $\alpha$. . . . . 98

FIGURE 10 : Plots of the asymptotic relative efficiency $\psi$ when the data set are evenly distributed, for different $\alpha$ and $\gamma$. See Theorem 3.4.4 for the properties of the ARE. . . . . 100

FIGURE 11 : Plots of optimal weights for different $\alpha$. . . . . 102

FIGURE 12 : Surface and contour plots of the optimal weights. . . . . 102

FIGURE 13 : Distributed risk as a function of the regularization parameter. We plot both the risk with optimal weights (MSE opt) and the risk obtained from sub-optimal averaging (MSE avg). We set $\alpha = 1$, $\gamma = 0.17$ and $k = 5, 10$. . . . . 103

FIGURE 14 : Limit of OE: (a) surface and (b) contour plots of $\mathcal{H}(\alpha^2, \gamma)$. . . . . 104

FIGURE 15 : Realized relative efficiency in a regression simulation. . . . . 112

FIGURE 16 : Distributed risk as a function of the regularization parameter. We plot the risk of the optimally weighted distributed estimator for an AR-1 covariance structure. We set $\alpha = 1$, $\gamma = 0.17$ and $k = 1, 2, 5, 10$. 114

FIGURE 17 : Million Song Year Prediction Dataset (MSD). Optimal weighted average (WONDER), Naive average, and regression on $1/k$ fraction of data. . . . . 115

FIGURE 18 : Illustration of Permutation PA for a rank-one signal in homogeneous noise. Permutations scramble signal structure, creating a noise-like matrix. PA selects those factors whose singular values rise above a chosen percentile of their permuted analogues, in this case correctly selecting one factor. . . . . 150

FIGURE 19 : Illustration of Permutation PA applied to a rank-one signal in *heterogeneous* noise. PA underestimates the noise level and dramatically over-selects as a consequence. . . . . 153

FIGURE 20 : Illustration of Signflip PA for the heterogeneous example of fig. 19, for which Permutation PA incorrectly selected nine components. Signflip PA accurately recovers the noise and correctly selects one component. . . . . 154

FIGURE 21 : Preview and rough intuition for theoretical analysis. Signflipping "destroys" low-rank signals (in operator norm) and consistently estimates the noise level—the singular values of signflipped data $R \circ X$ are close to those of the noise $N$. . . . . 157

FIGURE 22 : Regimes for delocalization rates $\mathbb{E}\|u\|_\infty, \mathbb{E}\|v\|_\infty = O(p^{-\alpha_1} \log^{-\alpha_2} p)$ for given signal strength rates $\theta = O(p^{\beta_1} \log^{\beta_2} p)$ from theorem 4.4.3: feasible range (north-east green), $L^1$ convergence (north-west blue and purple), and almost sure convergence (north-west blue). . . . 165

FIGURE 23 : The empirical and limiting spectral distributions of $R \circ N$ and $N_\pi$, where the first $n/2$ samples have entries with variance $1/(10n)$ and the remainder have entries with variance $9/(10n)$. Limiting spectral distributions shown on top using SPECTRODE (Dobriban, 2015). 174

FIGURE 24 : Mean rank selection $\hat{k}$ vs. signal strength $\theta$ for homogeneous noise of variance $1/n$. . . . . 182
FIGURE 25 : Heterogeneous noise where 90% of samples have noise variance $0.4/n$ and 10% of samples have noise variance $1/n$. . . . . 183
FIGURE 26 : Heterogeneous noise where 80% of features have noise variance $0.5/n$ on half the samples and $1.5/n$ on the second half, and the remaining features have noise variance $1/n$ on all samples. . . . . 184
FIGURE 27 : For heterogeneous noise added to realistically generated chlorine data, the components selected by signflip PA appear to rise above the noise and capture meaningful underlying latent behavior. On the other hand, permutations homogenize the noise and select many noise-like components. . . . . 185
FIGURE 28 : Empirical spectral distribution of the single-cell RNA-sequencing data with the cut-offs chosen by Permutation PA and Signflip PA (with both selection rules). . . . . 187
FIGURE 29 : Scatter plots of the first twelve left singular vectors of the scRNA-seq data. . . . . 188

# CHAPTER 1 : A Brief Introduction to Random Matrix Theory and Its Applications

Random Matrix Theory (RMT) traces back to the early days of statistical sciences in 1920s (Wishart) and the development of quantum mechanics in 1950s (Wigner). In quantum mechanics, the energy levels of a quantum system are described by eigenvalues of a Hermitian operator on a Hilbert space. Since the operator is infinite-dimensional, it is common to approximate the system by discretization. Hence, the limiting behavior of large dimensional random matrices has attracted special interest among physicists working in quantum mechanics. For more work on applications of RMT in physics, one can refer to Mehta (2004).

Statistics has entered into a new age where an increasingly larger volume of more complex data is being generated everyday. This brings the so-called high-dimensional data that are frequently associated with new phenomena beyond the boundary of classical multivariate statistics. Hence, RMT has emerged as a particularly useful framework and mathematical tool for formulating and answering many theoretical questions associated with the analysis of modern high-dimensional data. We will not spend too much time and effort on introducing rigorous definitions and mathematical details of RMT, since there are already many good references including review papers like Johnstone (2007); Paul and Aue (2014) and textbooks like Bai and Silverstein (2010); Anderson et al. (2010); Yao et al. (2015).

For our purpose, we will focus on "Marchenko-Pastur" (MP) type sample covariance matrices, which are fundamental and popular in statistics (see e.g., Bai and Silverstein (2010); Anderson (2003); Paul and Aue (2014); Yao et al. (2015)). A key concept is the spectral distribution, which for a $p \times p$ symmetric matrix $A$ is the distribution $F_A$ that places equal mass on all eigenvalues $\lambda_i(A)$ of $\Sigma$. This has cumulative distribution function (CDF) $F_A(x) = p^{-1} \sum_{i=1}^{p} \mathbf{1}(\lambda_i(A) \leq x)$. A central result in the area is the Marchenko-Pastur theorem, which states that eigenvalue distributions of sample covariance matrices converge (Marchenko and Pastur, 1967; Bai and Silverstein, 2010). We state the required assumptions below:

**Assumption 1.** *Consider the following conditions:*

1. *The $n \times p$ design matrix $X$ is generated as $X = Z\Sigma^{1/2}$ for an $n \times p$ matrix $Z$ with i.i.d. entries (viewed as coming from an infinite array), satisfying $\mathbb{E}[Z_{ij}] = 0$ and $\mathbb{E}[Z_{ij}^2] = 1$, and a deterministic $p \times p$ positive semidefinite population covariance matrix $\Sigma$.*

2. *The sample size $n$ grows to infinity proportionally with the dimension $p$, i.e. $n, p \to \infty$ and $p/n \to \gamma \in (0, \infty)$.*

3. *The sequence of spectral distributions $F_\Sigma := F_{\Sigma,n,p}$ of $\Sigma := \Sigma_{n,p}$ converges weakly to a limiting distribution $H$ supported on $[0, \infty)$, called the population spectral distribution.*

Then, the Marchenko-Pastur theorem states that with probability 1, the spectral distribution $F_{\widehat{\Sigma}}$ of the sample covariance matrix $\widehat{\Sigma} = X^\top X/n$ also converges weakly (in distribution) to a limiting distribution $F_\gamma := F_\gamma(H)$ supported on $[0, \infty)$ (Marchenko and Pastur, 1967; Bai and Silverstein, 2010). The limiting distribution is determined uniquely by a fixed-point equation for its *Stieltjes transform*, which is defined for any distribution $G$ supported on $[0, \infty)$ as

$$m_G(z) := \int_0^\infty \frac{1}{t - z} dG(t), \quad z \in \mathbb{C} \setminus \mathbb{R}^+.$$

With this notation, the Stieltjes transform of the spectral measure of $\widehat{\Sigma}$ satisfies

$$m_{\widehat{\Sigma}}(z) = p^{-1} \operatorname{tr}[(\widehat{\Sigma} - zI_p)^{-1}] \to_{a.s.} m_{F_\gamma}(z), \quad z \in \mathbb{C} \setminus \mathbb{R}^+,$$

where $m_{F_\gamma}(z)$ is the Stieltjes transform of $F_\gamma$. From the above brief introduction, one can already see the power of RMT: if the empirical quantities in a problem only depend on the eigenvalues of the data matrices, then we may probably apply RMT to study the problem and give a very precise characterization of the limiting behavior of the empirical quantities.

Now, we list several classical application areas of RMT in statistics. These include problems in dimension reduction, hypothesis testing, clustering, regression analysis and covariance estimation. For dimension reduction, RMT has been used to study PCA under the spiked

covariance model, which is first introduced by Johnstone (2001). This model has been studied extensively in the context of high-dimensional PCA since it brings a number of key issues associated with dimension reduction in the high-dimensional regime. See Johnstone and Paul (2018) for a wonderful and comprehensive survey of this area. People also studied similar models for CCA and MANOVA (Johnstone, 2008, 2009), where they gave an extensive account of the use of Roy's largest root test with Tracy-Widom limit distribution under the various classical double Wishart problems. In the domain of hypothesis testing, Bai et al. (2013) proposed a modification to Wilk's test in a high-dimensional setting based on a CLT for linear statistic of F-matrices. Bai et al. (2009) considered the test for equality of covariance matrices. They proposed corrections to the classical likelihood ratio test statistic when the number of features is proportional to the sample size. Lopes et al. (2011) proposed a test using averaged Hotelling's $T^2$ statistics based on random projections of the data into lower dimensional subspaces. Many statistical problems involve ridge-type shrinkage. A detailed study of such shrinkage has been carried out by El Karoui and Kösters (2011). Random matrices have also been applied to characterize properties of large random graphs, especially the limiting behavior of eigenvalues of graph Laplacian and the adjacency matrix (Ding and Jiang, 2010; Jiang, 2012). For more applications, one can refer to Paul and Aue (2014), we only discussed some of them due to space limitations.

In recent years, RMT also has been applied to solve modern machine learning problems. We will only list a few of them here. In the area of random projections and sketching, Dobriban and Liu (2019) considered sketching in high-dimensional least square regression. Later, sketching and cross-validation for high-dimensional ridge regression was also studied in Liu and Dobriban (2019). More recently, the performance of iterative Hessian sketch for least-squares problems was considered in Lacotte et al. (2020). In another popular research area, the mathematical theory for deep learning, RMT also has made significant contributions. Pennington and Worah (2017) developed a non-linear random matrix theory by extending the classical moments method, which is particularly useful in the theoretical study of neural networks. Hastie et al. (2019) studied minimum $\ell_2$ norm interpolation in high-dimensional

linear models and two-layer neural networks. Mei and Montanari (2019) used RMT to precisely capture the well-known double descent phenomenon in deep learning.

Later in the following chapters, we will extend the applications of RMT to some new areas like distributed machine learning and statistical analysis for heterogeneous data.

# CHAPTER 2 : Distributed Linear Regression by Averaging

This chapter is based on Dobriban and Sheng (2018), which is a joint work with my advisor Professor Edgar Dobriban. I contributed to a large portion of ideas, derivations, and simulations.

## 2.1. Introduction

Datasets are constantly increasing in size and complexity. This leads to important challenges for practitioners. Statistical inference and machine learning, which used to be computationally convenient on small datasets, now bring an enormous computational burden.

*Distributed computation* is a universal approach to deal with large datasets. Datasets are partitioned across several machines (or workers). The machines perform computations locally and communicate only small bits of information with each other. They coordinate to compute the desired quantity. This is the standard approach taken at large technology companies, which routinely deal with huge datasets spread over computing units. What are the best ways to divide up and coordinate the work?

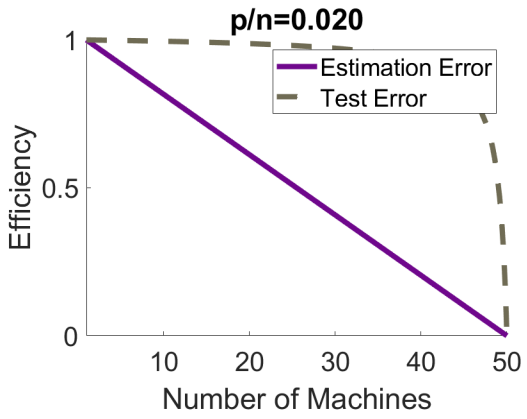



Figure 1: How much accuracy do we lose in distributed regression? The plots show the relative efficiency, i.e., the ratio of errors, of the global least squares (OLS) estimator, compared to the distributed estimator averaging the local least squares estimators. This efficiency is at most unity, because the global estimator is more accurate. If the efficiency is close to unity, then averaging is accurate. We show the behavior of estimation and test error, as a function of number of machines. We see that estimation error is *much more affected* than test error. The specific formulas are given in Table 1.

The same problem arises when the data is distributed due to privacy, security, or ethical concerns. For instance, medical and healthcare data is typically distributed across hospitals or medical units. The parties agree that they want to aggregate the results. At the same time, they do not want other parties access their data. How can they compute the desired aggregates, without sharing the data?

In both cases, the key question is how to do statistical estimation and machine learning in a distributed setting. And what performance can the best methods achieve? This is a question of broad interest, and it is expected that the area of distributed estimation and computation will grow even more in the future.

In this paper, we develop precise theoretical answers to fundamental questions in distributed estimation. We study *one-step and iterative parameter averaging* in statistical *linear models* under *data parallelism*. Specifically, suppose in the simplest case that we do linear regression (Ordinary Least Squares, OLS) on each subset of a dataset distributed over $k$ machines, and take an optimal weighted average of the regression coefficients. How do the statistical and predictive properties of this estimator compare to doing OLS on the full data?

We study the behavior of several learning and inference problems, such as *estimation error*, *test error* (i.e., out-of-sample *prediction error*), and *confidence intervals*. We also consider a high-dimensional (or proportional-limit) setting where the number of parameters is of the same order as the number of total samples (i.e., the size of the training data). We also study an analogous iterative algorithm, where we do local ridge regressions, take averages of the parameters on a central machine, send back the update to the local machines, and then again do local ridge, but where the penalty is centered around the previous mean. Our iterative algorithm falls between several classical methods such as ADMM and DANE, and we discuss connections.

We discover the following key phenomena, some of which are surprising in the context of existing work:

1. **Sub-optimality.** One-step averaging is not optimal (even with optimal weights), meaning that it leads to a performance decay. In contrast to some recent work (see the related work section), we find that there is a clear performance loss due to one-step averaging *even if we split the data only into two subsets*. This loss is because the number of parameters is of the same order as the sample size. However, we can quantify this loss precisely.

2. **Strong problem-dependence.** Different learning and inference problems are affected differently by the distributed framework. Specifically, *estimation error and the length of confidence intervals increases a lot, while prediction error increases less*. The intuition is that prediction is a noisy task, and hence the extra error incurred is relatively smaller.

3. **Simple form and universality.** The asymptotic efficiencies for one step distributed learning have simple forms that are often *universal*. Specifically, they do not depend on the covariance matrix of the data, or on the sample sizes on the local machines. For instance, the estimation efficiency *decreases linearly in the number of machines $k$* (see Figure 1 and Table 1).

4. **Iterative parameter averaging has benefits.** We show that simple iterative parameter averaging mechanisms can reduce the error efficiently. We also exhibit computation-statistics tradeoffs: depending on the hyperparameters, we can converge fast to statistically suboptimal solutions; or vice versa.

While there is already a lot of work in this direction (see Section 2.2) our results are new and complementary. The key elements of novelty of our setting are: (1) The sample size and the dimension are comparable, and we do not assume sparsity. (2) We have a new mathematical approach, using recent results from asymptotic random matrix theory such as (Rubio and Mestre, 2011). Our approach also develops a novel theoretical tool, the *calculus of deterministic equivalents*, and we illustrate how it can be useful in other problems as well. (3) We consider several accuracy metrics (estimation, prediction) in a unified framework of

Table 1: Estimation, Confidence Interval, and test efficiency as a function of number of machines $k$, the sample size $n$, and the dimension $p$. This is how much smaller the error of the global estimator is compared to the distributed estimator. These functions are plotted and described in Figure 1.

| Quantity | Relative efficiency $(n, p, k)$ |
|---|---|
| Estimation & CIs | $\frac{n-kp}{n-p}$ |
| Test error | $\frac{1}{1+\frac{p^2(k-1)}{n(n-kp)}}$ |

so-called general linear functionals.

### 2.2. Some related work

In this section we discuss some related work. There is a great deal of work in computer science and optimization on parallel and distributed computation (see e.g., Bertsekas and Tsitsiklis, 1989; Boyd et al., 2011; Bekkerman et al., 2011). In addition, there are several popular examples of distributed data processing frameworks: for instance MapReduce (Dean and Ghemawat, 2008) and Spark (Zaharia et al., 2010).

In contrast, there is less work on understanding the statistical properties, and the inherent computation-statistics tradeoffs, in distributed computation environments. This area has attracted increasing attention only in recent years, see for instance Mcdonald et al. (2009); Zhang et al. (2012, 2013b,a); Duchi et al. (2014); Zhang et al. (2015); Braverman et al. (2016); Jordan et al. (2016); Rosenblatt and Nadler (2016); Smith et al. (2016); Fan et al. (2017); Lin et al. (2017); Lee et al. (2017); Battey et al. (2018); Zhu and Lafferty (2018), and the references therein. See Huo and Cao (2018) for a review. We can only discuss the most closely related papers due to space limitations.

Zinkevich et al. (2009) study the parallelization of SGD for learning, by reducing it to the study of delayed SGD; giving positive results for low latency "multicore" settings. They give an insightful discussion of the impact of various computational platforms, such as shared memory architectures, clusters, and grid computing. Mcdonald et al. (2009) propose

averaging methods for special conditional maximum entropy models, showing variance reduction properties. Zinkevich et al. (2010) expand on this, proposing "parallel SGD" to average the SGD iterates computed on random subsets of the data. Their proof is based on the contraction properties of SGD.

Zhang et al. (2013b) bound the leading order term for MSE of averaged estimation in empirical risk minimization. Their bounds do not explicitly take dimension into account. However, their empirical data example clearly has large dimension $p$, considering a logistic regression with sample size $n = 2.4 \cdot 10^8$, and $p = 740,000$, so that $n/p \approx 340$. In their experiments, they distribute the data over up to 128 machines. So, our regime, where $k$ is of the same order as $n/p$, matches well their simulation setup. In addition, their concern is on regularized estimators, where they propose to estimate and reduce bias by subsampling.

Liu and Ihler (2014) study distributed estimation in statistical exponential families, connecting the efficiency loss from the global setting to the deviation from full exponential families. They also propose nonlinear KL-divergence-based combination methods, which can be more efficient than linear averaging.

Zhang et al. (2015) study divide and conquer kernel ridge regression, showing that the partition-based estimator achieves the statistical minimax rate over all estimators. Due to their generality, their results are more involved, and also their dimension is fixed. Lin et al. (2017) improve those results. Duchi et al. (2014) derive minimax bounds on distributed estimation where the number of bits communicated is controlled.

Rosenblatt and Nadler (2016) consider the distributed learning problem in three different settings. The first two settings are fixed dimensional. The third setting is high-dimensional M-estimation, where they study the first order behavior of estimators using prior results from Donoho and Montanari (2013); El Karoui et al. (2013). This is possibly the most closely related work to ours in the literature. They use the following representation, derived in the previous works mentioned above: a high-dimensional $M$-estimator can be written as

$\hat{\beta} = \beta + r(\gamma)\Sigma^{-1/2}\zeta(1 + o_P(1))$, where $\zeta \sim \mathcal{N}(0, I_p/p)$, $\gamma$ is the limit of $p/n$, and $r(\gamma)$ is a constant depending on the loss function, whose expression can be found in Donoho and Montanari (2013); El Karoui et al. (2013).

They derive a relative efficiency formula in this setting, which for OLS takes the form

$$\frac{\mathbb{E}\|\hat{\beta}_{dist} - \beta\|^2}{\mathbb{E}\|\hat{\beta} - \beta\|^2} = 1 + \gamma(1 - 1/k) + O(\gamma^2).$$

In contrast, our result for this case (Theorem 2.5.1) is equal to

$$\frac{1-\gamma}{1-k\gamma} = 1 + \gamma\frac{k-1}{1-k\gamma}.$$

Thus, our result is much more precise, and in fact exact, while of course being limited to the special case of linear regression.

In a heterogeneous data setting, Zhao et al. (2016) fit partially linear models, and estimate the common part by averaging. For model selection problems in GLM, Chen and Xie (2014) propose weighted majority voting methods. Lee et al. (2017) study sparse linear regression, showing that averaging debiased lasso estimators can achieve the optimal estimation rate if the number of machines is not too large. Battey et al. (2018) study a similar problem, also including hypothesis testing under more general sparse models. Shi et al. (2018); Banerjee et al. (2019b) show that in problems with non-standard rates, averaging can lead to improved pointwise inference, while decreasing performance in a uniform sense. Volgushev et al. (2019b) (among other contributions) provide conditions under which averaging quantile regression estimators have an optimal rate. Banerjee and Durot (2018) propose improvements based on communicating smoothed data, and fitting estimators after. Szabo and van Zanten (2018) study estimation methods under communication constraints in nonparametric random design regression model, deriving both minimax lower bounds and optimal methods.

See Section 2.7 for more discussion of multi-round methods.

## 2.3. One-step weighted averaging: General linear functionals

We consider the standard linear model

$$Y = X\beta + \varepsilon.$$

Here we have an outcome variable $y$ along with some $p$ covariates $x = (x^1, \ldots, x^p)^\top$, and want to understand their relationship. We observe $n$ such data points, arranging their outcomes into the $n \times 1$ vector $Y$, and their covariates into the $n \times p$ matrix $X$. We assume that $Y$ depends linearly on $X$, via some unknown $p \times 1$ parameter vector $\beta$.

We assume there are more samples than training data points, i.e., $n > p$, while $p$ can also be large. In that case, a gold standard is the usual least squares estimator (ordinary least squares or OLS)

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y.$$

We also assume that the coordinates of the noise $\varepsilon$ are uncorrelated and have variance $\sigma^2$.

Suppose now that the samples are distributed across $k$ machines (these can be real machines, but they can also be—say—sites or hospitals in medical applications, or mobile devices in federated learning). The $i$-th machine has the $n_i \times p$ matrix $X_i$, containing $n_i$ samples, and also the $n_i \times 1$ vector $Y_i$ of the corresponding outcomes for those samples. Thus, the $i$-th worker has access to only a subset of training $n_i$ data points out of the total of $n$ training data points. For instance, if the data points denote $n$ users, then they may be partitioned into $k$ sets based on country of residence, and we may have $n_1$ samples from the United States on one server, $n_2$ samples from Canada on another server, etc. The broad question is: How can we estimate the unknown regression parameter $\beta$ if we need to do most of the computations locally?

Let us write the partitioned data as

$$X = \begin{bmatrix} X_1 \\ \dots \\ X_k \end{bmatrix}, \; Y = \begin{bmatrix} Y_1 \\ \dots \\ Y_k \end{bmatrix}.$$

We also assume that each *local* OLS estimator $\hat{\beta}_i = (X_i^\top X_i)^{-1} X_i^\top Y_i$ is well defined, which requires that the number of local training data points $n_i$ must be at least $p$ on each machine (so $n_i \geq p$). We first consider combining the local OLS estimators at a parameter server via one-step weighted averaging. Since they are uncorrelated and unbiased for $\beta$, we consider unbiased weighted estimators

$$\hat{\beta}_{dist}(w) = \sum_{i=1}^{k} w_i \hat{\beta}_i$$

with $\sum_{i=1}^{k} w_i = 1$.

We introduce a "general linear functional" framework to study learning tasks such as estimation and prediction in a unified way. In the general framework, we predict *linear functionals* of $\beta$ of the form

$$L_A = A\beta + Z.$$

Here $A$ is a fixed $d \times p$ matrix, and $Z$ is a zero-mean Gaussian noise vector of dimension $d$, with covariance matrix $\mathrm{Cov}[Z] = h\sigma^2 I_d$, for some scalar parameter $h \geq 0$. We denote the covariance matrix between $\varepsilon$ and $Z$ by $N$, so that $\mathrm{Cov}[\varepsilon, Z] = N$. If $h = 0$, we say that there is no noise. In that case, we necessarily have $N = 0$.

We predict the linear functional $L_A$ via plug-in based on some estimator $\hat{\beta}_0$ (typically OLS or distributed OLS)

$$\hat{L}_A(\hat{\beta}_0) = A\hat{\beta}_0.$$

Table 2: A general framework for finite-sample efficiency calculations. The rows show the various statistical problems studied in our work, namely estimation, confidence interval formation, in-sample prediction, out-of-sample prediction and regression function estimation. The elements of the row show how these tasks fall in the framework of linear functional prediction described in the main body.

| Statistical learning problem | $L_A$ | $\hat{L}_A$ | $A$ | $h$ | $N$ |
|---|---|---|---|---|---|
| Estimation | $\beta$ | $\hat{\beta}$ | $I_p$ | 0 | 0 |
| Regression function estimation | $X\beta$ | $X\hat{\beta}$ | $X$ | 0 | 0 |
| Confidence interval | $\beta_j$ | $\hat{\beta}_j$ | $E_j^\top$ | 0 | 0 |
| Test error | $x_t^\top \beta + \varepsilon_t$ | $x_t^\top \hat{\beta}$ | $x_t^\top$ | 1 | 0 |
| Training error | $X\beta + \varepsilon$ | $X\hat{\beta}$ | $X$ | 1 | $\sigma^2 I_n$ |

We measure the quality of estimation by the mean squared error

$$M(\hat{\beta}_0) = \mathbb{E}\|L_A - \hat{L}_A(\hat{\beta}_0)\|^2.$$

We compute the *relative efficiency* of OLS $\hat{\beta}$ compared to a weighted distributed estimator $\hat{\beta}_{dist} = \hat{\beta}_{dist}(w)$:

$$E(A, d; X_1, \ldots, X_k) := \frac{M(\hat{\beta})}{M(\hat{\beta}_{dist})}.$$

The relative efficiency is a fundamental quantity, giving the loss of accuracy due to distributed estimation.

*2.3.1. Examples*

We now show how several learning and inference problems fall into the general framework. See Table 2 for a concise summary.

- **Parameter estimation**. In parameter estimation, we want to estimate the regression coefficient vector $\beta$ using $\hat{\beta}$. This is an example of the general framework by taking $A = I_p$, and without noise (so that $h = 0$).

- **Regression function estimation**. We can use $X\hat{\beta}$ to estimate the regression function $\mathbb{E}(Y|X) = X\beta$. In this case, the transform matrix is $A = X$, the linear functional is $L_A = X\beta$, the predictor is $\hat{L}_A = X\hat{\beta}$, and there is no noise.

- **Out-of-sample prediction (Test error)**. For out-of-sample prediction, or test error, we consider a test data point $(x_t, y_t)$, generated from the same model $y_t = x_t^\top \beta + \varepsilon_t$, where $x_t, \varepsilon_t$ are independent of $X, \varepsilon$, and only $x_t$ is observable. We want to use $x_t^\top \hat{\beta}$ to predict $y_t$.

  This corresponds to predicting the linear functional $L_{x_t} = x_t^\top \beta + \varepsilon_t$, so that $A = x_t^\top$, and the noise is $Z = \varepsilon_t$, which is uncorrelated with the noise $\varepsilon$ in the original problem.

- **In-sample prediction (Training error)**. For in-sample prediction, or training error, we consider predicting the response vector $Y$, using the model fit $X\hat{\beta}$. Therefore, the functional $L_A$ is $L_A = Y = X\beta + \varepsilon$. This agrees with regression function estimation, except for the noise $Z = \varepsilon$, which is identical to the original noise. Hence, the noise scale is $h = 1$, and $N = \mathrm{Cov}\,[\varepsilon, Z] = \sigma^2 I_n$.

- **Confidence intervals**. To construct confidence intervals for individual coordinates, we consider the normal model $Y \sim \mathcal{N}(X\beta, \sigma^2 I_n)$. Assuming $\sigma^2$ is known, a confidence interval with coverage $1 - \alpha$ for a given coordinate $\beta_j$ is

$$\hat{\beta}_j \pm \sigma z_{\alpha/2} V_j^{1/2},$$

  where $z_\alpha = \Phi^{-1}(\alpha)$ is the inverse normal CDF, and $V_j$ is the $j$-th diagonal entry of $(X^\top X)^{-1}$.

  Therefore, we can measure the difficulty of the problem by $V_j$. The larger $V_j$ is, the longer the confidence interval. This also measures the difficulty of estimating the coordinate $L_A = \beta_j$. This can be fit in our general framework by choosing $A = E_j^\top$, the $1 \times p$ vector of zeros, with only a one in the $j$-th coordinate. This problem is noiseless. In this sense,

the problem of confidence intervals is the same as the estimation accuracy for individual coordinates of $\beta$.

If $\sigma$ is not known, then we we first need to estimate it in a distributed way. This is an interesting problem in itself, but beyond the scope of our current work.

*2.3.2. Finite sample results*

We now show how to calculate the efficiency explicitly in the general framework. We start with the simpler case where $h = 0$. We then have for the OLS estimator

$$M(\hat{\beta}) = \sigma^2 \cdot \operatorname{tr}\left[(X^\top X)^{-1} A^\top A\right].$$

For the distributed estimator with weights $w_i$ summing to one, given by $\hat{\beta}_{dist}(w) = \sum_i w_i \hat{\beta}_i$, we have

$$M(\hat{\beta}_{dist}) = \sigma^2 \cdot \left(\sum_{i=1}^{k} w_i^2 \cdot \operatorname{tr}\left[(X_i^\top X_i)^{-1} A^\top A\right]\right).$$

Using a simple Cauchy-Schwarz inequality (see Section 2.8.1 for the argument for parameter estimation), we find that the optimal efficiency, for the optimal weights, is

$$E(A; X_1, \ldots, X_k) = \operatorname{tr}\left[(X^\top X)^{-1} A^\top A\right] \cdot \sum_{i=1}^{k} \frac{1}{\operatorname{tr}\left[(X_i^\top X_i)^{-1} A^\top A\right]}. \tag{2.1}$$

This shows that the key to understanding the efficiency are the traces $\operatorname{tr}\left[(X_i^\top X_i)^{-1} A^\top A\right]$. Proving that the efficiency is at most unity turns out to require the concavity of the matrix functional $1/\operatorname{tr}(X^{-1} A^\top A)$. This is a consequence of classical results in convex analysis, see for instance Davis (1957); Lewis (1996). For completeness, we give a short self-contained proof in Section 2.8.2 of the Appendix.

**Proposition 2.3.1** (Concavity for general efficiency, Davis (1957); Lewis (1996))**.** *The function $f(X) = 1/\operatorname{tr}(X^{-1} A^\top A)$ is a concave function defined on positive definite matrices. As a consequence, the general relative efficiency for distributed estimation is at most unity*

*for any matrices $X_i$:*

$$E(A; X_1, \ldots, X_k) \leq 1.$$

For the more general case when $h \neq 0$, we can also find the OLS MSE as

$$M(\hat{\beta}) = \sigma^2 \cdot \left[ \operatorname{tr}\left( (X^\top X)^{-1} A^\top A \right) - 2 \operatorname{tr}\left( A(X^\top X)^{-1} X^\top N \right) + hd \right].$$

For the distributed estimator, we can find, denoting $N_i := \operatorname{Cov}[\varepsilon_i, Z]$,

$$M(\hat{\beta}_{dist}) = \sigma^2 \cdot \left( \sum_{i=1}^{k} w_i^2 \cdot \operatorname{tr}\left[ (X_i^\top X_i)^{-1} A^\top A \right] - 2 w_i \cdot \operatorname{tr}\left( A(X_i^\top X_i)^{-1} X_i^\top N_i \right) \right) + \sigma^2 hd.$$

Let $a_i = \operatorname{tr}\left[ (X_i^\top X_i)^{-1} A^\top A \right]$, and $b_i = \operatorname{tr}\left( A(X_i^\top X_i)^{-1} X_i^\top N_i \right)$. The optimal weights can be found from a quadratic optimization problem:

$$w_i = \frac{\lambda^* + b_i}{a_i}, \ \lambda^* := \frac{1 - \sum_{i=1}^{k} \frac{b_i}{a_i}}{\sum_{i=1}^{k} \frac{1}{a_i}}.$$

The resulting formula for the optimal weights, and for the global optimum, can be calculated explicitly. The details can be found in the supplement (Section 2.8.3).

## 2.4. Calculus of deterministic equivalents

### *2.4.1. A calculus of deterministic equivalents in RMT*

We saw that the relative efficiency depends on the trace functionals $\operatorname{tr}[(X^\top X)^{-1} \ A^\top A]$, for specific matrices $A$. To find their limits, we will use the technique of *deterministic equivalents* from random matrix theory. This is a method to find the almost sure limits of random quantities. See for example Hachem et al. (2007); Couillet et al. (2011) and the related work section below.

For instance, the well known Marchenko-Pastur (MP) law for the eigenvalues of random matrices (Marchenko and Pastur, 1967; Bai and Silverstein, 2010) states that the eigenvalue distribution of certain random matrices is asymptotically deterministic. More generally,

one of the best ways to understand the MP law is that *resolvents are asymptotically deterministic*. Indeed, let $\widehat{\Sigma} = n^{-1}X^\top X$, where $X = Z\Sigma^{1/2}$ and $Z$ is a random matrix with iid entries of zero mean and unit variance. Then the MP law means that for any $z$ with positive imaginary part, we have the equivalence

$$(\widehat{\Sigma} - zI)^{-1} \asymp (x_p\Sigma - zI)^{-1},$$

for a certain scalar $x_p = x(\Sigma, n, p, z)$ (that will be specified later). At this stage we can think of the equivalence entry-wise, but we will make this precise next. The above formulation has appeared in some early works by VI Serdobolskii, see e.g., Serdobolskii (1983), and Theorem 1 on page 15 of Serdobolskii (2007) for a very clear statement.

The consequence is that simple linear functionals of the random matrix $(\widehat{\Sigma} - zI)^{-1}$ have a deterministic equivalent based on $(x_p\Sigma - zI)^{-1}$. In particular, we can approximate the needed trace functionals by simpler deterministic quantities. For this we will take a principled approach and define some appropriate notions for a *calculus of deterministic equivalents*, which allows us to do calculations in a simple and effective way.

First, we make more precise the notion of equivalence. We say that the (deterministic or random) not necessarily symmetric matrix sequences $A_n, B_n$ of growing dimensions are *equivalent*, and write

$$A_n \asymp B_n$$

if

$$\lim_{n\to\infty} |\mathrm{tr}\,[C_n(A_n - B_n)]| = 0$$

almost surely, for any sequence $C_n$ of not necessarily symmetric matrices with bounded trace norm, i.e., such that

$$\limsup \|C_n\|_{tr} < \infty.$$

We call such a sequence $C_n$ a *standard sequence*. Recall here that the trace norm (or nuclear

norm) is defined by $\|M\|_{tr} = \mathrm{tr}((M^\top M)^{1/2}) = \sum_i \sigma_i$, where $\sigma_i$ are the singular values of $M$.

*2.4.2. General MP theorem*

To find the limits of the efficiencies, the most important deterministic equivalent will be the following result, essentially a consequence of the generalized Marchenko-Pastur theorem of Rubio and Mestre (2011) (see Section 2.8.4 for the argument). We study the more general setting of elliptical data. In this model the data samples may have different scalings, having the form $x_i = g_i^{1/2}\Sigma^{1/2}z_i$, for some vector $z_i$ with iid entries, and for datapoint-specific *scale parameters* $g_i$. Arranging the data as the rows of the matrix $X$, that takes the form

$$X = \Gamma^{1/2} Z \Sigma^{1/2},$$

where $Z$ and $\Gamma$ are as before: $Z$ has iid standardized entries, while $\Sigma$ is the covariance matrix of the features. Now $\Gamma$ is the diagonal *scaling matrix* containing the scales $g_i$ of the samples. This model has a long history in multivariate statistics (e.g., Mardia et al., 1979).

**Theorem 2.4.1** (Deterministic equivalent in elliptical models, consequence of Rubio and Mestre (2011))**.** *Let the $n \times p$ data matrix $X$ follow the elliptical model*

$$X = \Gamma^{1/2} Z \Sigma^{1/2},$$

*where $\Gamma$ is an $n \times n$ diagonal matrix with non-negative entries representing the scales of the $n$ observations, and $\Sigma$ is a $p \times p$ positive definite matrix representing the covariance matrix of the $p$ features. Assume the following:*

1. *The entries of $Z$ are iid random variables with mean zero, unit variance, and finite $8 + c$-th moment, for some $c > 0$.*
2. *The eigenvalues of $\Sigma$, and the entries of $\Gamma$, are uniformly bounded away from zero and infinity.*
3. *We have $n, p \to \infty$, with $\gamma_p = p/n$ bounded away from zero and infinity.*

*Let $\widehat{\Sigma} = n^{-1}X^\top X$ be the sample covariance matrix. Then $\widehat{\Sigma}$ is equivalent to a scaled version of the population covariance*

$$\widehat{\Sigma}^{-1} \asymp \Sigma^{-1} \cdot e_p.$$

*Here $e_p = e_p(n, p, \Gamma) > 0$ is the unique solution of the fixed-point equation*

$$1 = \frac{1}{n} \operatorname{tr}\left[e_p \Gamma (I + \gamma_p e_p \Gamma)^{-1}\right].$$

Thus, the inverse sample covariance matrix has a deterministic equivalent in terms of a scaled version of the inverse population covariance matrix. This result does not require the convergence of the aspect ratio $p/n$, or of the e.s.d. of $\Sigma$, and $\Gamma$, as is sometimes the case in random matrix theory. However, if the empirical spectral distribution of the scales $\Gamma$ tends to $G$, the above equation has the limit

$$\int \frac{se}{1 + \gamma se} dG(s) = 1.$$

The usual MP theorem is a special case of the above result where $\Gamma = I_n$. As a result, we obtain the following corollary:

**Corollary 2.4.2** (Deterministic equivalent in MP models)**.** *Let the $n \times p$ data matrix $X$ follow the model $X = Z\Sigma^{1/2}$, where $\Sigma$ is a $p \times p$ positive definite matrix representing the covariance matrix of the $p$ features. Assume the same conditions on $\Sigma$ from Theorem 2.4.1. Then $\widehat{\Sigma}$ is equivalent to a scaled version of the population covariance*

$$\widehat{\Sigma}^{-1} \asymp \frac{1}{1 - \gamma_p} \cdot \Sigma^{-1}.$$

The proof is immediate, by checking that $e_p = 1/(1 - \gamma_p)$ in this case.

*2.4.3. Related work on deterministic equivalents*

There are several works in random matrix theory on deterministic equivalents. One of the early works is Serdobolskii (1983), see Serdobolskii (2007) for a modern summary. The name "deterministic equivalents" and technique was more recently introduced and re-popularized by Hachem et al. (2007) for signal-plus-noise matrices. Later Couillet et al. (2011) developed deterministic equivalents for matrix models of the type $\sum_{k=1}^{B} R_k^{1/2} X_k T_k X_k^\top R_k^{1/2}$, motivated by wireless communications. See the book Couillet and Debbah (2011) for a summary of related work. See also Müller and Debbah (2016) for a tutorial. However, many of these results are stated only for some fixed functional of the resolvent, such as the Stieltjes transform. One of our points here is that there is a much more general picture.

Rubio and Mestre (2011) is one of the few works that explicitly states more general convergence of arbitrary trace functionals of the resolvent. Our results are essentially a consequence of theirs.

However, we think that it is valuable to define a set of rules, a "calculus" for working with deterministic equivalents, and we use those techniques in our paper. Similar ideas for operations on deterministic equivalents have appeared in Peacock et al. (2008), for the specific case of a matrix product. Our approach is more general, and allows many more matrix operations, see below.

*2.4.4. Rules of calculus*

The calculus of deterministic equivalents has several properties that simplify calculations. We think these justify the name of *calculus*. Below, we will denote by $A_n, B_n, C_n$ etc, sequences of deterministic or random matrices. See Section 2.8.5 in the supplement for the proof.

**Theorem 2.4.3** (Rules of calculus)**.** *The calculus of deterministic equivalents has the following properties.*

1. **Equivalence.** *The $\asymp$ relation is indeed an equivalence relation.*

2. **Sum.** *If* $A_n \asymp B_n$ *and* $C_n \asymp D_n$*, then* $A_n + C_n \asymp B_n + D_n$*.*

3. **Product.** *If* $A_n$ *is a sequence of matrices with bounded operator norms i.e.,* $\|A_n\|_{op} < \infty$*, and* $B_n \asymp C_n$*, then* $A_n B_n \asymp A_n C_n$*.*

4. **Trace.** *If* $A_n \asymp B_n$*, then* $\operatorname{tr}\{n^{-1}A_n\} - \operatorname{tr}\{n^{-1}B_n\} \to 0$ *almost surely.*

5. **Stieltjes transforms.** *As a consequence, if* $(A_n - zI_n)^{-1} \asymp (B_n - zI_n)^{-1}$ *for symmetric matrices* $A_n, B_n$*, then* $m_{A_n}(z) - m_{B_n}(z) \to 0$ *almost surely. Here* $m_{X_n}(z) = n^{-1}\operatorname{tr}(X_n - zI_n)^{-1}$ *is the Stieltjes transform of the empirical spectral distribution of* $X_n$*.*

In addition, the calculus of deterministic equivalents has additional properties, such as continuous mapping theorems, differentiability, etc. We have developed the differentiability in the follow-up work (Dobriban and Sheng, 2019).

We also briefly sketch several applications of the calculus of deterministic equivalents in Section 2.8.6 in the supplement, to studying the risk of ridge regression in high dimensions, including in the distributed setting, gradient flow for least squares, interpolation in high dimensions, heteroskedastic PCA, as well as exponential family PCA. We emphasize that in each case, including for the formulas of asymptotic efficiencies in the current work, there are other proof techniques, but they tend to be more case-by-case. The calculus provides a unified set of methods, and separate results can be seen as applications of the same approach.

## 2.5. Examples

We now use the calculus of deterministic equivalents to find the limits of the trace functionals in our general framework. We study each problem in turn. For asymptotics, we consider as before elliptical models. The data on the $i$-th machine takes the form

$$X_i = \Gamma_i^{1/2} Z_i \Sigma^{1/2},$$

where $\Gamma_i$ contains the *scales* of the $i$-th machine and $Z_i$ is the appropriate submatrix of $X$.

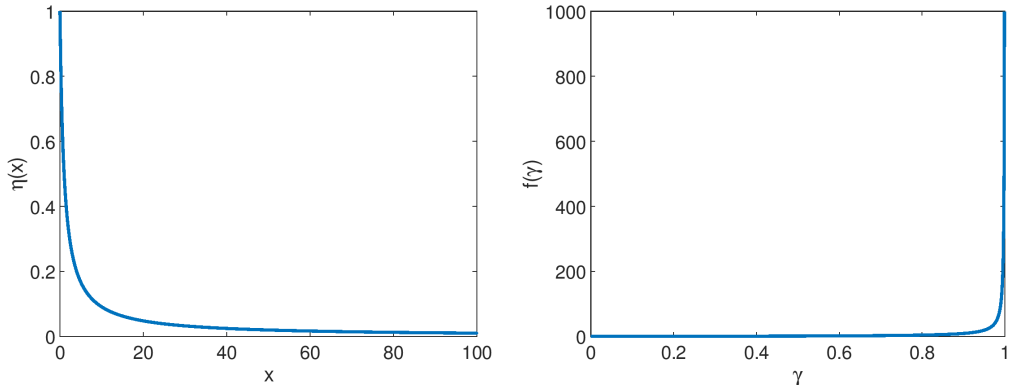



Figure 2: Plots of $\eta$ and $f$ for $G$ being the point mass at unity.

In this model, it turns out that the efficiencies can be expressed in a simple way via the $\eta$-transform (Tulino and Verdú, 2004). The $\eta$-transform of a distribution $G$ is

$$\eta(x) = \mathbb{E}_G \frac{1}{1+xT},$$

for all $x$ for which this expectation is well-defined. We will see that the efficiencies can be expressed in terms of the functional inverse $f$ of the $\eta$-transform evaluated at the specific value $1-\gamma$:

$$f(\gamma, G) = \eta_G^{-1}(1-\gamma). \tag{2.2}$$

We think of elliptical models where the limiting distribution of the scales $g_1, \ldots, g_n$ is $G$. For some insight on the behavior of $\eta$ and $f$, consider first the case when $G$ is a point mass at unity, $G = \delta_1$. In this case, all scales are equal, so this is just the usual Marchenko-Pastur model. Then, we have $\eta(x) = 1/(1+x)$, while $f(\gamma, G) = \gamma/(1-\gamma)$. See Figure 2 for the plots. The key points to notice are that $\eta$ is a decreasing function of $x$, with $\eta(0) = 1$, and $\lim_{x\to\infty} \eta(x) = 0$. Moreover, $f$ is an increasing function on $[0,1]$ with $f(0) = 0$, $\lim_{\eta\to 1} f(\eta) = +\infty$. The same qualitative properties hold in general for compactly supported distributions $G$ bounded away from 0.

*2.5.1. Parameter estimation*

For estimating the parameter, we have $\mathbb{E}\|\beta - \hat{\beta}\|^2 = \sigma^2 \operatorname{tr}(X^\top X)^{-1}$. We find via (2.1) the estimation efficiency

$$RE(X_1, \ldots, X_k) = \operatorname{tr}[(X^\top X)^{-1}] \cdot \left[\sum_{i=1}^{k} \frac{1}{\operatorname{tr}[(X_i^\top X_i)^{-1}]}\right].$$

Recall that $X^\top X = \sum_{i=1}^{k} X_i^\top X_i$. Recall that the empirical spectral distribution (e.s.d.) of a symmetric matrix $M$ is simply the CDF of its eigenvalues (which are all real-valued). More formally, it is the discrete distribution $F_p$ that places equal mass on all eigenvalues of $M$.

**Theorem 2.5.1** (RE for elliptical and MP models)**.** *Under the conditions of Theorem 2.4.1, suppose that, as $n_i \to \infty$ with $p/n_i \to \gamma_i \in (0, 1)$, the e.s.d. of $\Gamma$ converges weakly to some $G$, the e.s.d. of each $\Gamma_i$ converges weakly to some $G_i$, and that the e.s.d. of $\Sigma$ converges weakly to $H$. Suppose that $H$ is compactly supported away from the origin, while $G$ is also compactly supported and does not have a point mass at the origin. Then, the RE has almost sure limit*

$$ARE = f(\gamma, G) \cdot \sum_{i=1}^{k} \frac{1}{f(\gamma_i, G_i)}.$$

*For Marchenko-Pastur models, the RE has the form $(1/\gamma - k)/(1/\gamma - 1)$.*

See Section 2.8.7 in the supplement for the proof. For MP models, for any finite sample size $n$, dimension $p$, and number of machines $k$, we can approximate the ARE as

$$ARE \approx \frac{n - kp}{n - p}.$$

This efficiency for MP models depends on a simple linear way on $k$. We find this to be a surprisingly simple formula, which can also be easily computed in practice. Moreover, the formula has several more intriguing properties:

1. The ARE *decreases linearly* with the number of machines $k$. This holds as long as

$ARE \geq 0$. At the threshold case $ARE = 0$, there is a phase transition. The reason is that there is a singularity, and the OLS estimator is undefined for at least one machine.

However, we should be cautious about interpreting the linear decrease. For the root mean squared error (RMSE), the efficiency is the square root of the ARE above, and thus does not have a linear decrease.

2. The ARE has two important *universality* properties.

   (a) First, it *does not depend* on how the samples are distributed across the different machines, i.e., it is independent of the specific sample sizes $n_i$.

   (b) Second, it *does not depend* on the covariance matrix $\Sigma$ of the samples. This is in contrast to the estimation error of OLS, which does in fact depend on the covariance structure. Therefore, we think that the cancellation of $\Sigma$ in the ARE is noteworthy.

The ARE is also very accurate in simulations. See Figure 3 for an example. Here we report the results of a simulation where we generate an $n \times p$ random matrix $X$ such that the rows are distributed independently as $x_i \sim \mathcal{N}(0, \Sigma)$. We take $\Sigma$ to be diagonal with entries chosen uniformly at random between 1 and 2. We choose $n > p$, and for each value of $k$ such that $k < n/p$, we split the data into $k$ groups of a random size $n_i$. To ensure that each group has a size $n_i \geq p$, we first let $n_i^0 = p$, and then distribute the remaining samples uniformly at random. We then show the theoretical results compared to the theoretical ARE. We observe that the two agree closely.

*2.5.2. Regression function estimation*

For estimating the regression function, we have $\mathbb{E}\|X(\beta - \hat{\beta})\|^2 = \sigma^2 p$. We then find via equation (2.1) the prediction efficiency

$$FE(X_1, \ldots, X_k) = \sum_{i=1}^{k} \frac{p}{\mathrm{tr}((X_i^\top X_i)^{-1} X^\top X))}.$$

For asymptotics, we consider as before elliptical models.

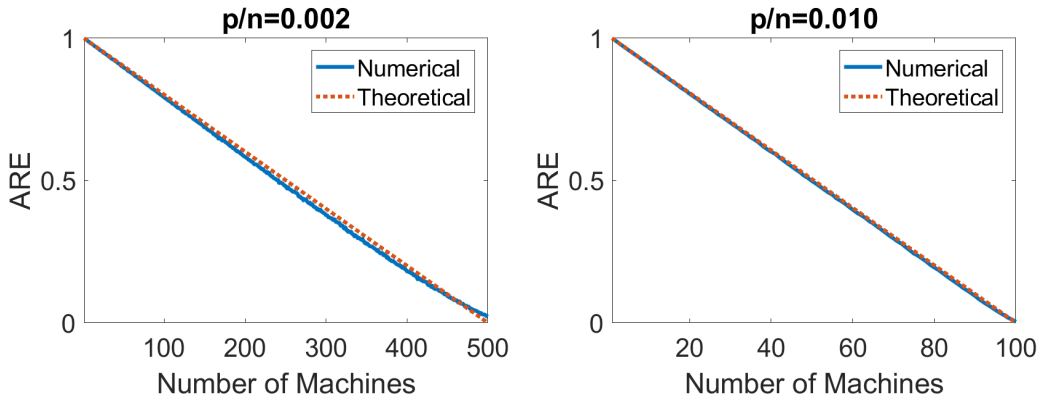



Figure 3: Comparison of empirical and theoretical ARE for standard sample covariance matrices. Left: $n = 10,000$, $p = 20$. Right: $n = 10,000$, $p = 100$.

**Theorem 2.5.2** (FE for elliptical and MP models)**.** *Under the conditions of Theorems 2.4.1 and 2.5.1, the FE has the almost sure limit*

$$FE(X_1, \ldots, X_k) \to_{a.s.} \sum_{i=1}^{k} \frac{1}{1 + \left(\frac{1}{\gamma}\mathbb{E}_G T - \frac{1}{\gamma_i}\mathbb{E}_{G_i} T\right) f(\gamma_i, G_i)}.$$

*Under Marchenko-Pastur models, the conditions of Corollary 2.4.2, the FE has the almost sure limit* $\frac{\gamma}{1-\gamma}\sum_{i=1}^{k}\frac{1-\gamma_i}{\gamma_i}$.

See Section 2.8.10 for the proof. This efficiency is more complex than that for estimation error; specifically it generally depends on the individual $\gamma_i$ and not just $\gamma$.

*2.5.3. In-sample prediction (Training error)*

For in-sample prediction, we start with the well known formula

$$\mathbb{E}||X(\beta - \hat{\beta}) + \varepsilon||^2 = \sigma^2[n - \mathrm{tr}((X^\top X)^{-1} X^\top X] = \sigma^2(n - p).$$

As we saw, to fit in-sample prediction in the general framework, we need to take the transform matrix $A = X$, the noise $Z = \varepsilon$, and the covariance matrices $N_i = \mathrm{Cov}\,[\varepsilon_i, Z] = \mathrm{Cov}\,[\varepsilon_i, \varepsilon]$. Then, in the formula for optimal weights we need to take $a_i = \mathrm{tr}[(X_i^\top X_i)^{-1} X^\top X]$ and $b_i = \mathrm{tr}(X(X_i^\top X_i)^{-1} X_i^\top N_i) = \mathrm{tr}[(X_i^\top X_i)^{-1} X_i^\top N_i X] = \mathrm{tr}[(X_i^\top X_i)^{-1} X_i^\top X_i] = p$. Therefore,

the optimal error for distributed regression is achieved by the weights

$$w_i = \frac{\lambda - b_i}{a_i} = \frac{\lambda - p}{a_i}, \quad \lambda = \frac{1 - \sum_{i=1}^k \frac{b_i}{a_i}}{\sum_{i=1}^k \frac{1}{a_i}} = \frac{1}{\sum_{i=1}^k \frac{1}{a_i}} - p.$$

Plugging these into $M(\hat{\beta}_{dist})$ given in the general framework, we find

$$M(\hat{\beta}_{dist}) = \sigma^2 \left( n - 2p + \frac{1}{\sum_{i=1}^k \frac{1}{a_i}} \right), \quad a_i = \mathrm{tr}((X_i^\top X_i)^{-1} X^\top X).$$

Thus, the optimal in-sample prediction efficiency is

$$IE(X_1, \ldots, X_k) = \frac{n-p}{n - 2p + \frac{1}{\sum_{i=1}^k \frac{1}{\mathrm{tr}((X_i^\top X_i)^{-1} X^\top X)}}}.$$

For asymptotics in elliptical models, we find:

**Theorem 2.5.3** (IE for elliptical and MP models)**.** *Under the conditions of Theorems 2.4.1 and 2.5.1, the IE has the almost sure limit*

$$IE(X_1, \ldots, X_k) \to_{a.s.} \frac{1-\gamma}{1 - 2\gamma + \frac{1}{\sum_{i=1}^k \psi(\gamma_i, G_i)}},$$

*where $\psi$ is the following functional of the distributions $G_i$ and $G$, depending on the inverse of the $\eta$-transform $f$ defined in equation* (2.2)*:*

$$\psi(\gamma_i, G_i) = \frac{1}{\gamma + (\mathbb{E}_G T - \frac{\gamma}{\gamma_i} \mathbb{E}_{G_i} T) f(\gamma_i, G_i)}.$$

*Under the conditions of Corollary 2.4.2, the IE has the almost sure limit*

$$IE(X_1, \ldots, X_k) \to_{a.s.} \frac{1-\gamma}{1 - 2\gamma + \frac{\gamma(1-\gamma)}{1-k\gamma}} = \frac{1}{1 + \frac{(k-1)\gamma^2}{(1-k\gamma)(1-\gamma)}}.$$

See Section 2.8.11 for the proof. This efficiency does not depend on a simple linear way on

$k$, but rather via a ratio of two linear functions of $k$. However, it can be checked that many of the properties (e.g., monotonicity) for ARE still hold here.

*2.5.4. Out-of-sample prediction (Test error)*

In out-of-sample prediction, we consider a test datapoint $(x_t, y_t)$, generated from the same model $y_t = x_t^\top \beta + \varepsilon_t$, where $x_t, \varepsilon_t$ are independent of $X, \varepsilon$, and only $x_t$ is observable. We want to use $x_t^\top \hat{\beta}$ to predict $y_t$. We compare the prediction error of two estimators:

$$OE(x_t; X_1, \ldots, X_k) := \frac{\mathbb{E}\left[(y_t - x_t^\top \hat{\beta})^2\right]}{\mathbb{E}\left[(y_t - x_t^\top \hat{\beta}_{dist})^2\right]}.$$

In our general framework, we saw that this corresponds to predicting the linear functional $x_t^\top \beta + \varepsilon_t$. Based on equation (2.1), the optimal out-of-sample prediction efficiency is

$$OE(x_t; X_1, \ldots, X_k) = \frac{1 + x_t^\top (X^\top X)^{-1} x_t}{1 + \frac{1}{\sum_{i=1}^k \frac{1}{x_t^\top (X_i^\top X_i)^{-1} x_t}}}.$$

For asymptotics in elliptical models, we find the following result. Since the samples have the form $x_i = g_i^{1/2} \Sigma^{1/2} z_i$, the test sample depends on a scale parameter $g_t$.

**Theorem 2.5.4** (OE for elliptical and MP models)**.** *Under the conditions of Theorems 2.4.1 and 2.5.1, the OE has the almost sure limit, conditional on $g_t$*

$$OE(x_t; X_1, \ldots, X_k) \to_{a.s.} \frac{1 + g_t \cdot f(\gamma, G)}{1 + \frac{g_t}{\sum_{i=1}^k \frac{1}{f(\gamma_i, G_i)}}}.$$

*For Marchenko-Pastur models under the conditions of Corollary 2.4.2, the OE has the almost sure limit*

$$\frac{\frac{1}{1-\gamma}}{1 + \frac{\gamma}{1-k\gamma}} = \frac{1}{1 + \frac{(k-1)\gamma^2}{1-k\gamma}}.$$

See Section 2.8.12 for the proof. If the scale parameter $g_t$ is random, then the OE typically does not have an almost sure limit, and converges in distribution to a random variable

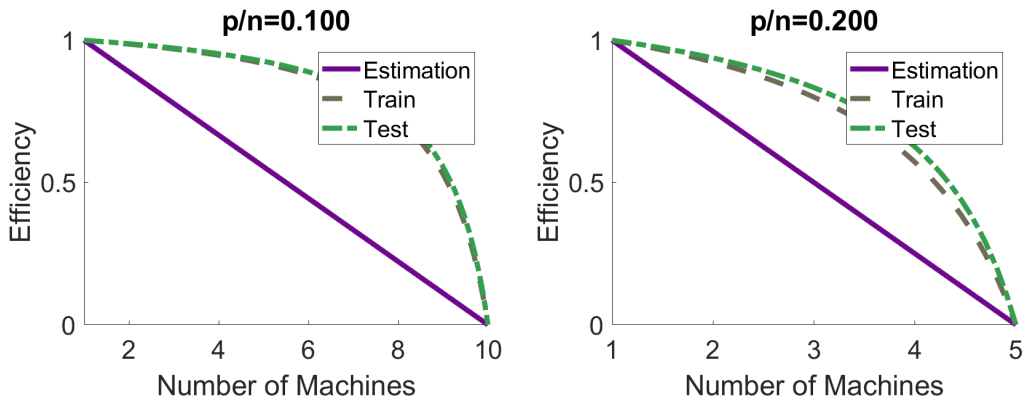



Figure 4: Relative efficiency for Marchenko-Pastur model.

instead. We mention that Theorem 2.5.4 holds under even weaker conditions, if we are only given the $4 + c$-th moment of $z_1$ instead of $8 + c$-th one. The argument is slightly different, and is presented in the location referenced above.

One can check that that $OE \geq RE$. Thus, out-of-sample prediction incurs a smaller efficiency loss than estimation. The intuition is that the out-of-sample prediction always involves a fixed loss due to the *irreducible noise* in the test sample, which "amortizes" the problem. Moreover,

$$OE \geq IE \geq RE.$$

The intuition here is that IE incurs a smaller fixed loss than OE, because the noise in the training set is effectively reduced, as it is already partly fit by our estimation process. So the graph of IE will be in between the other two criteria. See Figure 4. We also see that the IE is typically very close to OE.

In addition, the increase of the *reducible* part of the error is the same as for estimation error. The prediction error has two components: the irreducible noise, and the reducible error. The reducible error has the same behavior as for estimation, and thus on figure 4 it would have the same plot as the curve for estimation.

*2.5.5. Confidence intervals*

To form confidence intervals, we consider the normal model $Y \sim \mathcal{N}(X\beta, \sigma^2 I_n)$. Recall that in this model the OLS estimator has distribution $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2(X^\top X)^{-1})$. Assuming $\sigma^2$ is known, an exact level $1-\alpha$ confidence interval for a given coordinate $\beta_j$ can be formed as

$$\hat{\beta}_j \pm \sigma z_{\alpha/2} V_j^{1/2},$$

where $z_\alpha = \Phi^{-1}(\alpha)$ is the inverse normal CDF, and $V_j$ is the $j$-th diagonal entry of $(X^\top X)^{-1}$. We follow the same program as before, comparing the length of the confidence intervals formed based on our two estimators. However, for technical reasons it is more convenient to work with squared length.

Thus we consider the criterion

$$CE(j; X_1, \ldots, X_k) := \frac{V_j}{V_{j,dist}}.$$

Here $V_{j,dist}$ is the variance of the $j$-th entry of an optimally weighted distributed estimator. As we saw in our framework, this is equivalent to estimating the $j$-th coordinate of $\beta$. Hence the optimal confidence interval efficiency is

$$CE(j; X_1, \ldots, X_k) = [(X^\top X)^{-1}]_{jj} \cdot \sum_{i=1}^{k} \frac{1}{[(X_i^\top X_i)^{-1}]_{jj}}. \tag{2.3}$$

For asymptotics, we find:

**Theorem 2.5.5** (CE for elliptical and MP models)**.** *Under the conditions of Theorems 2.4.1 and 2.5.1, the CE has the same limit as the ARE from Theorem 2.5.1. Therefore, for Marchenko-Pastur models, the CE also has the form before, $CE(j) = (1/\gamma - k)/(1/\gamma - 1)$.*

See Section 2.8.13 for the proof.

*2.5.6. Understanding and comparing the efficiencies*

We give two perspectives for understanding and comparing the efficiencies. The key qualitative insight is that estimation and CIs are much more affected than prediction.

**Criticality of $k$.** We ask: What is the largest number of machines such that the asymptotic efficiency is at least 1/2? Let us call this the *critical* number of machines. It is easy to check that for estimation and CIs, $k_R = (\gamma + 1)/(2\gamma)$. For training error, $k_{Tr} = (\gamma^2 - \gamma + 1)/\gamma$, while for test error, $k_{Te} = (\gamma^2 + 1)/(\gamma^2 + \gamma)$.

We also have the following asymptotics as $\gamma \to 0$:

$$k_R \asymp 1/(2\gamma),$$

while

$$k_{Tr} \asymp k_{Te} \asymp 1/\gamma.$$

So the number of machines that can be used is nearly maximal (i.e., $n/p$) for training and test error, while it is about *half that* for estimation error and CIs. This shows quantitatively that estimation and CIs are much more affected by distributed averaging than prediction.

**Edge efficiency.** The maximum number of machines that we can use is approximately $k^* = 1/\gamma - 1$, for small $\gamma$. Let us define the *edge efficiency* $e^*$ as the relative efficiency achieved at this edge case. For estimation and CIs, we have $e^*_R = \gamma/(1 - \gamma)$. For training error, $e^*_{Tr} = (1 - \gamma)/(2 - 3\gamma)$, and for test error, $e^*_{Te} = 1/[2(1 - \gamma)]$.

We also have the following asymptotic values as $\gamma \to 0$:

$$e^*_R \asymp \gamma,$$

while

$$e^*_{Tr} \asymp \frac{1}{2} + \frac{\gamma}{4}, \text{ and } e^*_{Te} \asymp \frac{1}{2} + \frac{\gamma}{2}.$$

This shows that for $n \gg p$ the edge efficiency is vanishing for estimation and CIs, while it is approximately 1/2 for training and test error. Thus, even for the maximal number of machines, prediction error is not greatly increased.

## 2.6. Insights for Parameter Estimation

There are additional insights for the special case of parameter estimation. First, it is of interest to understand the performance of one-step weighted averaging with suboptimal weights $w_i$. How much do we lose compared to the optimal performance if we do not use the right weights? In practice, it may seem reasonable to take a simple average of all estimators. We have performed that analysis in the supplement (Section 2.8.14), and we found that the loss can be viewed in terms of an inequality between the arithmetic and harmonic means.

There are several more remarkable properties. We have studied the monotonicity properties and interpretation of the relative efficiency, see the supplement (Section 2.8.15). We have also given a multi-response regression characterization that heuristically gives an upper bound on the "*degrees of freedom*" for distributed regression (Section 2.8.16).

For elliptical data, the graph of ARE is a curve below the straight line from before. The interpretation is that for elliptical distributions, there is a larger efficiency loss in one-step weighted averaging. Intuitively, the problem becomes "more non-orthogonal" due to the additional variability from sample to sample.

## 2.7. Multi-shot methods

While our focus has been on methods with one round of communication, in practice it is more common to use iterative methods with several rounds of communication. These usually improve statistical accuracy. A great deal of research has been done on multi-shot distributed algorithms. Due to limited space, here we will only list and analyze some of them. Our least squares objective can be written as a sum of least squares objectives for each machine as

$$f(\beta) = \frac{1}{k}\sum_{i=1}^{k} f_i(\beta) = \frac{1}{k}\sum_{i=1}^{k} \|X_i\beta - Y_i\|_2^2.$$

Here each machine has access only to local data $(X_i, Y_i)$. With this formulation, there are a large number of standard optimization methods to minimize this objective: distributed gradient descent, alternating directions method of multipliers, and several others we discuss below. We will focus on parameter server architectures, where each machine communicates independently with a central server.

**Distributed Gradient Descent.** A simple multi-round approach to distributed learning is synchronous distributed gradient descent (DGD), as discussed e.g., in Chu et al. (2007). This maintains iterates $\hat{\beta}^t$, started with some standard value, such as $\hat{\beta}^0 = 0$. At each iteration $t$ each local machine calculates the gradient $\nabla f_i(\hat{\beta}^t)$ at the current iterate $\hat{\beta}^t$, and then sends the local gradient to the server to obtain the overall gradient

$$\nabla f(\hat{\beta}^t) = \frac{1}{k} \sum_{i=1}^{k} \nabla f_i(\hat{\beta}^t).$$

Then the center server sends the updated parameter $\hat{\beta}^{t+1} = \hat{\beta}^t - \alpha \nabla f(\hat{\beta}^t)$ back to the local machines, where $\alpha$ is the learning rate (LR). This synchronous implementation is *identical* to centralized gradient descent. Thus for smooth and strongly convex objectives and suitably small $\alpha$, $\mathcal{O}(L/\lambda \log(1/\varepsilon))$ communication rounds are sufficient to attain an $\varepsilon$-suboptimal solution in terms of objective value, where $L, \lambda$ are the smoothness and strong convexity parameters (e.g., Boyd and Vandenberghe, 2004).

1. Many works study the optimization properties of GD/synchronous DGD, in terms of convergence rate to the optimal objective or parameter value. From a statistical point of view, the GD iterates start with large bias and small variance, and gradually reduce bias, while slightly increasing the variance. This has motivated work on the risk properties of GD, emphasizing early stopping (Yao et al., 2007; Ali et al., 2019, e.g.,). Recently, Ali et al. (2019) gave a more refined analysis of the estimation risk of GD for OLS, showing that its risk at an optimal stopping time is at most 1.22 times the risk of optimally tuned ridge regression.

2. Compared to GD, one-shot weighted averaging has several advantages: it is simpler to implement, as it requires no iterations. It requires fewer tuning parameters, and those can be set optimally in an easy way, unlike the LR $\alpha$. The weights are proportional to $1/\operatorname{tr}[(X_i^T X_i)^{-1}]$, which can be computed locally. We point out that GD is sensitive to the learning rate: this has to be bounded (by $2/\lambda_{\max}(X^\top X)$ for OLS) to converge, and the convergence can be faster for large LR, hence in practice sophisticated LR schedules are used. This can make DGD complicated to use. In addition, in practice DGD is susceptible to stragglers, i.e., machines that take too long to compute their answers. To mitigate this problems, asynchronous DGD algorithms (e.g., Tsitsiklis et al., 1986; Nedic and Ozdaglar, 2009), and other sophisticated coding ideas (Tandon et al., 2017) have been proposed. However those lead to additional complexity and hyperparameters to tune (e.g., for async algorithms: how much to wait, how to aggregate non-straggler gradients).

3. One may also use other gradient based methods, such as accelerated or quasi-Newton methods, e.g., L-BFGS (Agarwal et al., 2014).

**Alternating Direction Method of Multipliers (ADMM).** Another approach is the alternating direction method of multipliers (see Boyd et al., 2011, for an exposition) and its variants. In ADMM, we alternate between solving local problems, global averaging, and computing local dual variables. For us, at time step $t$ of ADMM, each local machine calculates a local estimator

$$\hat{\beta}_i^{t+1} = (X_i^\top X_i + \rho I)^{-1}[X_i^\top Y_i + \rho(\hat{\beta}^t - u_i^t)],$$

(where $\rho$ is a hyperparameter) and sends it to the parameter server to get an average

$$\hat{\beta}^{t+1} = \frac{1}{k}\sum_{i=1}^{k} \hat{\beta}_i^{t+1}.$$

Finally, the server sends $\hat{\beta}^{t+1}$ back to the local machines to update the dual variables

$$u_i^{t+1} = u_i^t + \hat{\beta}_i^{t+1} - \hat{\beta}^{t+1}.$$

These three steps can be written as a linear recursion $z^{t+1} = Az^t + b$ for a state variable $z^t$ including $\hat{\beta}^t, \hat{\beta}_i^t$ and $u_i^t$. If all singular values of $A$ are less than one, then the iteration converges to a fixed point solving $z = Az + b$, so that $z = (I - A)^{-1}b$. However, it seems hard to prove convergence in our asymptotic setting.

**Distributed Approximate Newton-type Method (DANE).** Shamir et al. (2014) proposed an approximate Newton-like method (DANE), which uses that the sub-problems are similar. For our problem, DANE aggregates the local gradients on the parameter server at each step $t$, and sends this quantity, i.e., $X^\top(X\hat{\beta}^t - Y)/(2k)$ to all machines. Then each machine computes a local estimator by a gradient step in the direction of a regularized local Hessian $X_i^\top X_i + \rho I$,

$$\hat{\beta}_i^{t+1} = \hat{\beta}^t + \frac{\eta}{k} \cdot (X_i^\top X_i + \rho I)^{-1} X^\top (Y - X\hat{\beta}^t),$$

where $\rho$ is the regularizer and $\eta$ is the learning rate. The machines send it to the server to get the aggregated estimator

$$\hat{\beta}^{t+1} = \frac{1}{k} \sum_{i=1}^{k} \hat{\beta}_i^{t+1}.$$

For a noiseless model where $Y = X\beta$, we can summarize the update rule as

$$\hat{\beta}^{t+1} - \beta = \left( I - \frac{\eta}{k^2} \cdot \sum_{i=1}^{k} \left( X_i^\top X_i + \rho I \right)^{-1} X^\top X \right) \left( \hat{\beta}^t - \beta \right),$$

so we have the error bound

$$\|\hat{\beta}^t - \beta\|_2 \leq \left\| I - \frac{\eta}{k^2} \cdot \sum_{i=1}^{k} \left( X_i^\top X_i + \rho I \right)^{-1} X^\top X \right\|_2^t \cdot \|\hat{\beta}^0 - \beta\|_2.$$

In Shamir et al. (2014), the authors showed that given a suitable learning rate $\eta$ and

regularizer $\rho$, when $X_i^\top X_i$ is close to $X^\top X/k$, $\hat{\beta}^t \to \beta$ as $t \to \infty$.

For a noisy linear model $Y = X\beta + \varepsilon$, the limit of $\hat{\beta}^t$ is exactly the OLS estimator of the whole data set, and we have the following recursion formula

$$\hat{\beta}^{t+1} - (X^\top X)^{-1}X^\top Y = \left(I - \frac{\eta}{k^2} \cdot \sum_{i=1}^{k} \left(X_i^\top X_i + \rho I\right)^{-1} X^\top X\right)\left(\hat{\beta}^t - (X^\top X)^{-1}X^\top Y\right),$$

and the convergence guarantee is the same as for the noiseless case.

**Iterative Averaging Method.** Here we describe an iterative averaging method for distributed linear regression. This method turns out to be connected to DANE, and it has the advantage that it can be analyzed more conveniently. We define a sequence of *local estimates* $\hat{\beta}_i^t$ and *global estimates* $\hat{\beta}^t$ with initialization $\hat{\beta}^0 = 0$. At the $t$-th step we update the local estimate by the following weighted average of the local ridge regression estimator and the current global estimate $\hat{\beta}^t$

$$\hat{\beta}_i^{t+1} = \left(X_i^\top X_i + n_i\rho_i I\right)^{-1}\left(X_i^\top Y_i + n_i\rho_i\hat{\beta}^t\right).$$

Then we average the local estimates

$$\hat{\beta}^{t+1} = \frac{1}{k}\sum_i \hat{\beta}_i^{t+1}.$$

To understand this, let us first consider a noiseless model where $Y_i = X_i\beta$. In that case, we can also write this update as a weighted average,

$$\hat{\beta}_i^{t+1} = (I - W_i)\beta + W_i\hat{\beta}^t,$$

where

$$W_i = n_i\rho_i \cdot \left(X_i^\top X_i + n_i\rho_i I\right)^{-1}$$

is the weight matrix of the global estimate. Propagating the iterative update to the global

machine, we find a linear update rule:

$$\hat{\beta}^{t+1} = \frac{1}{k}\sum_i W_i\hat{\beta}^t + \left(I - \frac{1}{k}\sum_i W_i\right)\beta = W\hat{\beta}^t + (I-W)\beta,$$

where $W = \frac{1}{k}\sum_i W_i$. Hence, the error is updated as

$$\hat{\beta}^{t+1} - \beta = W \cdot [\hat{\beta}^t - \beta] = \left(I - \frac{1}{k}\sum_{i=1}^{k}\left(X_i^\top X_i + n_i\rho_i I\right)^{-1} X_i^\top X_i\right)\left(\hat{\beta}^t - \beta\right).$$

This recursion relation is very similar to the one for DANE; we just need to replace $X^\top X/k$ by $X_i^\top X_i$ (and in practice usually $\eta = 1$ is used). The only difference is that DANE has a step where we need to collect the local gradients to get the global gradient, and then broadcast it to all local machines. Our iterative averaging method has lower communication cost.

In terms of convergence, $\hat{\beta}^{t+1}$ will converge geometrically to $\beta$ for all $\beta$, if and only if the largest eigenvalue of $W$ is strictly less than 1. It is not hard to see that this holds if at least one $X_i^\top X_i$ has positive eigenvalues by using the fact $\lambda_{\max}(A+B) \leq \lambda_{\max}(A) + \lambda_{\max}(B)$. When the samples are uniformly distributed, we should have $X^\top X/k \approx X_i^\top X_i$, which means the convergence rates of DANE and iterative averaging should be very close. Hence, in terms of the total cost (communication and computation), our iterative averaging should compare favorably to DANE.

To summarize the noiseless case, we can formulate the following result:

**Theorem 2.7.1** (Convergence of iterative averaging, noiseless case)**.** *Consider the iterative averaging method described above. In the noiseless case when $Y_i = X_i\beta$, we have the following: If at least one $X_i^\top X_i$ has positive eigenvalues, then the iterates converge to the true coefficients geometrically, $\hat{\beta}^t \to \beta$, and:*

$$\|\hat{\beta}^t - \beta\|_2 \leq \lambda_{\max}\left(\frac{1}{k}\sum_{i=1}^{k} n_i\rho_i \cdot \left(X_i^\top X_i + n_i\rho_i I\right)^{-1}\right)^t \cdot \|\beta\|_2.$$

Consider now the noisy case when $Y_i = X_i\beta + \varepsilon_i$ with the same assumptions as in the rest of the paper. We have

$$\begin{aligned}\hat{\beta}_i^{t+1} &= W_i\hat{\beta}^t + (I - W_i)\beta + \left(X_i^\top X_i + n_i\rho_i I\right)^{-1} X_i^\top \varepsilon_i \\ &=_d W_i\hat{\beta}^t + (I - W_i)\beta + \left(X_i^\top X_i + n_i\rho_i I\right)^{-1} X_i^\top X_i \cdot Z_i \\ &= W_i\hat{\beta}^t + (I - W_i)(\beta + Z_i),\end{aligned}$$

where $Z_i \sim \mathcal{N}(0, \sigma^2[X_i^\top X_i]^{-1})$. As before, defining $Z$ appropriately

$$\begin{aligned}\hat{\beta}^{t+1} &= W\hat{\beta}^t + (I - W)\beta + \frac{1}{k}\sum_{i=1}^{k}(I - W_i)Z_i \\ &= W\hat{\beta}^t + (I - W)\beta + Z,\end{aligned}$$

so $\hat{\beta}^{t+1} - \beta = W \cdot [\hat{\beta}^t - \beta] + Z$.

With noise, $\hat{\beta}^t$ does not converge to OLS, but to the following quantity:

$$\hat{\beta}_* = \left(\sum_{i=1}^{k}\left(X_i^\top X_i + n_i\rho_i I\right)^{-1} X_i^\top X_i\right)^{-1} \cdot \sum_{i=1}^{k}\left(X_i^\top X_i + n_i\rho_i I\right)^{-1} X_i^\top Y_i.$$

We can check that $\hat{\beta}_*$ is an unbiased estimator for $\beta$ and $\hat{\beta}^{t+1} - \hat{\beta}_* = W \cdot [\hat{\beta}^t - \hat{\beta}_*]$.

Under the conditions of Theorem 2.7.1, we have $\hat{\beta}^t \to \hat{\beta}_*$, and the MSE for $\hat{\beta}_*$ is

$$\begin{aligned}\mathbb{E}\|\hat{\beta}_* - \beta\|^2 = \mathbb{E}\|(I - W)^{-1}Z\|^2 &= \mathbb{E}\|\frac{1}{k}\sum_{i=1}^{k}(I - W)^{-1}(I - W_i)Z_i\|^2 \\ &= \frac{\sigma^2}{k^2}\sum_{i=1}^{k}\mathrm{tr}[(I - W)^{-2}(X_i^\top X_i + n_i\rho_i I)^{-2}X_i^\top X_i] \\ &= \sigma^2\sum_{i=1}^{k}\mathrm{tr}\left[\left(\sum_{i=1}^{k}\left(X_i^\top X_i + n_i\rho_i I\right)^{-1} X_i^\top X_i\right)^{-2}\left(X_i^\top X_i + n_i\rho_i I\right)^{-2} X_i^\top X_i\right].\end{aligned}$$

How large is this MSE, and how does it depend on $\rho_i$?We have the following results.

**Theorem 2.7.2** (Properties of Iterative averaging, noisy case)**.** *Consider the iterative averaging method described above. In the noisy case when $Y_i = X_i\beta + \varepsilon$, we have the following:*

1. *If at least one $X_i^\top X_i$ has strictly positive eigenvalues, then the iterates converge to the following limiting unbiased estimator*

$$\hat{\beta}_* = \left(\sum_{i=1}^k \left(X_i^\top X_i + n_i\rho_i I\right)^{-1} X_i^\top X_i\right)^{-1} \cdot \sum_{i=1}^k \left(X_i^\top X_i + n_i\rho_i I\right)^{-1} X_i^\top Y_i,$$

   *and the convergence is geometric*

$$\|\hat{\beta}^t - \hat{\beta}_*\|_2 \le \lambda_{\max}\left(\frac{1}{k}\sum_{i=1}^k n_i\rho_i \cdot \left(X_i^\top X_i + n_i\rho_i I\right)^{-1}\right)^t \cdot \|\hat{\beta}_*\|_2.$$

2. *The mean squared error of $\hat{\beta}_*$ has the following form*

$$\mathbb{E}\|\hat{\beta}_* - \beta\|^2 = \sigma^2 \sum_{i=1}^k \operatorname{tr}\left[\left(\sum_{i=1}^k \left(X_i^\top X_i + n_i\rho_i I\right)^{-1} X_i^\top X_i\right)^{-2} \left(X_i^\top X_i + n_i\rho_i I\right)^{-2} X_i^\top X_i\right].$$

3. *Suppose the samples are evenly distributed, i.e., $n_1 = n_2 = \cdots = n_k = n/k$ and the regularizers are all the same $\rho_1 = \rho_2 = \cdots = \rho_k = \rho$. The MSE is a differentiable function $\psi(\rho)$ of the regularizer $\rho \in [0, +\infty)$, with derivative*

$$\begin{aligned}\psi'(\rho) = \frac{2k}{n} \operatorname{tr}[\Delta^{-1} \sum_{i=1}^k \left(\widehat{\Sigma}_i + \rho I\right)^{-2} \widehat{\Sigma}_i \cdot \Delta^{-2} \sum_{i=1}^k \left(\widehat{\Sigma}_i + \rho I\right)^{-2} \widehat{\Sigma}_i \\ - \Delta^{-2} \sum_{i=1}^k \left(\widehat{\Sigma}_i + \rho I\right)^{-3} \widehat{\Sigma}_i],\end{aligned}$$

   *where $\widehat{\Sigma}_i = X_i^\top X_i/n_i$ and $\Delta := \sum_{i=1}^k \left(\widehat{\Sigma}_i + \rho I\right)^{-1} \widehat{\Sigma}_i$.*

4. *$\psi(\rho)$ is a non-increasing function on $[0, +\infty)$ and $\psi'(0) < 0$. So for any $\rho > 0, \psi(\rho) < \psi(0)$, i.e. the MSE of the iterative averaging estimator with positive regularizer is smaller than the MSE of the one-step averaging estimator.*

5. *When $\rho = 0$, $\hat{\beta}_*$ reduces to the one-step averaging estimator $1/k \cdot \sum_{i=1}^{k} (X_i^\top X_i)^{-1} X_i^\top Y_i$ with MSE*

$$\psi(0) = \sigma^2/k^2 \cdot \sum_{i=1}^{k} \operatorname{tr}(X_i^\top X_i)^{-1}.$$

*When $\rho \to +\infty$, $\hat{\beta}_*$ converges to the OLS estimator $(X^\top X)^{-1} X^\top Y$ with MSE*

$$\lim_{\rho \to +\infty} \psi(\rho) = \sigma^2 \operatorname{tr}(X^\top X)^{-1}.$$

See Section 2.8.17 of the supplement for the proof. The argument for monotonicity relies on Schur complements, and is quite nontrivial. From Theorem 2.7.2, it appears we should choose the regularizer $\rho$ as large as possible, since the limiting estimator $\hat{\beta}_*$ will converge to the OLS estimator as $\rho \to \infty$. This is true for statistical accuracy. However, there is a computational tradeoff, since the convergence rate of $\hat{\beta}^t$ to $\hat{\beta}_*$ is slower for large $\rho$.

Moreover, one may argue that $\hat{\beta}_*$ reduces to the naive averaging estimator but not the optimally weighted averaging estimator when $\rho = 0$. However, we have shown in the supplement (Section 2.8.14) that for evenly distributed samples, the MSE of the naive averaging estimator and the optimally weighted averaging estimator is asymptotically the same. Thus, there exists a regularizer such that the iterative averaging estimator has smaller MSE than the one-step weighted averaging estimator.

**Other approaches.** There are many other approaches to distributed learning. *Dual averaging for decentralized optimization* over a network (Duchi et al., 2011) builds on Nesterov's dual averaging method (Nesterov, 2009). It chooses the iterates to minimize an averaged first-order approximation to the function, regularized with a proximal function. The *communication-efficient surrogate likelihood* approximates the objective by an expression

of the form $\tilde{f}(\beta) = f_1(\beta) - \beta^\top(\nabla f_1(\bar{\beta}) - \nabla f(\bar{\beta}))$, where $\bar{\beta}$ is a preliminary estimator (Jordan et al., 2016; Wang et al., 2017). Chen et al. (2018b) propose a related method for quantile regression. Both are related to DANE (Shamir et al., 2014).

Chen et al. (2018a) study divide and conquer SGD (DC-SGD), running SGD on each machine and averaging the results. They also propose a distributed first-order Newton-type estimator starting with a preliminary estimator $\bar{\beta}$, of the form $\bar{\beta} - \Sigma^{-1}(k^{-1}\sum_i \nabla f_i(\bar{\beta}))$, where $\Sigma$ is the population Hessian. They show how to numerically estimate this efficiently, and also develop a more accurate multi-round version.

*2.7.1. Numerical comparisons*

We report simulations to compare the convergence rate and statistical accuracy of the one-shot weighted method with some popular multi-shot methods described above (Figure 5). Here we work with a linear model $Y = X\beta + \varepsilon$, where $X, \beta$ and $\varepsilon$ all follow standard normal distributions. We take $n = 10000, p = 100$, and $k = 20$. We plot the relative efficiencies of different methods against the number of iterations.

We can see that the one-shot weighted method is good in some cases. The multi-shot methods usually need several iterations to achieve better statistical accuracy. When the communication cost is large, one-shot methods are attractive. Also, we can clearly see the computation vs accuracy tradeoff for the iterative averaging method from the plots. When the regularizer is small, the convergence is fast, but in the end the accuracy is not as good as the other multi-shot methods. On the other hand, if the regularizer is large, we have a better accuracy with slower convergence. Moreover, the widely-used multi-shot methods can require a lot of work for parameter tuning, and sometimes it is very difficult to find the optimal parameters. In contrast, weighted averaging requires less tuning, making it a more attractive method.

We have performed several more numerical simulations to verify our theory, in addition to the results shown in the paper. Due to space limitations, these are presented in the supplement. In Section 2.8.18, we present an empirical data example to assess the accuracy

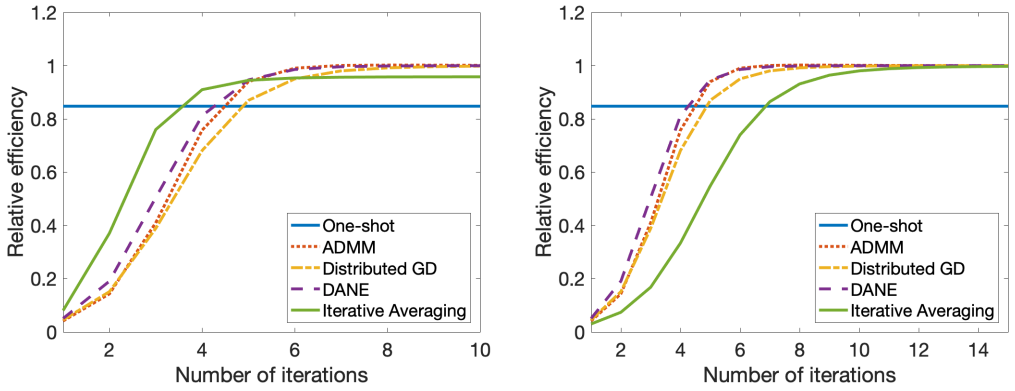



Figure 5: Comparison of the one-shot weighted method and several widely used multi-shot methods,

of our theoretical results for one-shot averaging. We find that they can be quite accurate.

## 2.8. Appendix

### *2.8.1. Form of optimal weights*

Here we describe the proof of the form of optimal weights for the special case of parameter estimation. Each local estimator is unbiased, and has MSE $M_i = \sigma^2 \operatorname{tr}[(X_i^\top X_i)^{-1}]$. If we restrict to $\sum_{i=1}^k w_i = 1$, then the weighted estimator is unbiased and its MSE equals

$$MSE(w) = \sum_i w_i^2 \cdot MSE(\hat{\beta}_i) = \sum_i w_i^2 \cdot M_i.$$

Clearly, to minimize this subject to $\sum_i w_i = 1$, by the Cauchy-Schwarz inequality we should take $w_i^* = M_i^{-1}/(\sum_j M_j^{-1})$, and the minimum is $1/(\sum_j M_j^{-1})$. This finishes the proof.

### *2.8.2. Proof of Proposition 2.3.1*

Notice that, it is sufficient to show that, for any given $A$, the function $f(X) = 1/\operatorname{tr}(X^{-1}A^\top A)$ is concave on positive definite matrices. Similar to the proof of proposition 2.2., we can define $g(t) = f(X + tV)$. The constraints on $X, V, X + tV$ are the same as before. Now, we have

$$
\begin{aligned}
g(t) &= \frac{1}{\mathrm{tr}((X+tV)^{-1}A^\top A)} = \frac{1}{\mathrm{tr}((I+tX^{-1/2}VX^{-1/2})^{-1}X^{-1/2}A^\top AX^{-1/2})} \\
&= \frac{1}{\mathrm{tr}((I+t\Lambda)^{-1}Q^\top X^{-1/2}A^\top AX^{-1/2}Q)} \\
&= \left(\sum_{i=1}^{n} \frac{(Q^\top X^{-1/2}A^\top AX^{-1/2}Q)_{ii}}{1+t\lambda_i}\right)^{-1}.
\end{aligned}
$$

Since $Q^\top X^{-1/2}A^\top AX^{-1/2}Q$ is always nonnegative definite, we can get the desired result by an explicit calculation. We show below the steps for $A = I$ for simplicity, but the same steps extend to general $A$.

Let us define $g(t) = f(X + tV)$, where $X \succ 0$ is a positive definite matrix and $V$ is any symmetric matrix such that $X + tV \succ 0$ is still positive definite. Then $f(X)$ is concave iff $g(t)$ is concave on its domain for any $X$ and $V$.

Now we have

$$
\begin{aligned}
g(t) = \frac{1}{\mathrm{tr}[(X+tV)^{-1}]} &= \frac{1}{\mathrm{tr}[X^{-1}(I+tX^{-1/2}VX^{-1/2})^{-1}]} \\
&= \frac{1}{\mathrm{tr}[X^{-1}Q(I+t\Lambda)^{-1}Q^\top]} \\
&= \frac{1}{\mathrm{tr}[Q^\top X^{-1}Q(I+t\Lambda)^{-1}]} \\
&= \left(\sum_{i=1}^{n} \frac{(Q^\top X^{-1}Q)_{ii}}{1+t\lambda_i}\right)^{-1},
\end{aligned}
$$

where $\lambda_i$-s are eigenvalues of $X^{-1/2}VX^{-1/2}$. From the assumption, we always have $1+t\lambda_i > 0$. Since $Q^\top X^{-1}Q$ is a positive definite matrix, its diagonal elements are all positive. We may use the notation $\alpha_i = (Q^\top X^{-1}Q)_{ii}$. Then, let us compute $g'(t)$ and $g''(t)$. First we have

$$
g'(t) = \left(\sum_{i=1}^{n} \frac{\alpha_i}{1+\lambda_i t}\right)^{-2} \cdot \left(\sum_{i=1}^{n} \frac{\alpha_i\lambda_i}{(1+\lambda_i t)^2}\right),
$$

Next, we find

$$g''(t) = 2\Big(\sum_{i=1}^n \frac{\alpha_i}{1+\lambda_i t}\Big)^{-3} \cdot \Big[\Big(\sum_{i=1}^n \frac{\alpha_i\lambda_i}{(1+\lambda_i t)^2}\Big)^2 - \Big(\sum_{i=1}^n \frac{\alpha_i}{1+\lambda_i t}\Big)\Big(\sum_{i=1}^n \frac{\alpha_i\lambda_i^2}{(1+\lambda_i t)^3}\Big)\Big]$$

Multiplying by $-2\Big(\sum_{i=1}^n \frac{\alpha_i}{1+\lambda_i t}\Big)^3$, we get the expression

$$\begin{aligned}
&\sum_{1\le i<j\le n} \frac{\alpha_i\alpha_j\lambda_j^2}{(1+\lambda_i t)(1+\lambda_j t)^3} + \frac{\alpha_j\alpha_i\lambda_i^2}{(1+\lambda_j t)(1+\lambda_i t)^3} - \frac{2\alpha_i\alpha_j\lambda_i\lambda_j}{(1+\lambda_i t)^2(1+\lambda_j t)^2} \\
&= \sum_{1\le i<j\le n} \frac{\alpha_i\alpha_j}{(1+\lambda_i t)^3(1+\lambda_j t)^3}[\lambda_j^2(1+\lambda_i t)^2 + \lambda_i^2(1+\lambda_j t)^2 - 2\lambda_i\lambda_j(1+\lambda_i t)(1+\lambda_j t)] \\
&= \sum_{1\le i<j\le n} \frac{\alpha_i\alpha_j(\lambda_i-\lambda_j)^2}{(1+\lambda_i t)^3(1+\lambda_j t)^3} \ge 0.
\end{aligned}$$

Hence $g(t)$ concave, and so is $f(X)$.

We can use the convexity directly to check $RE$ is less than or equal to unity. Indeed, $f$ is affine, in the sense that $f(cX) = cf(X)$ for any $c > 0$. The concavity result that we proved implies that, with $A_i = X_i^\top X_i$,

$$\sum_{i=1}^k f(A_i)/k \le f\left(\sum_{i=1}^k A_i/k\right).$$

By the affine nature of $f$, this result implies that $f$ is sub-additive. This can be checked to be equivalent to $RE \le 1$, finishing the proof.

*2.8.3. Computing optimal weights in the general framework, Section 2.3.1*

Recall that we have

$$\begin{aligned}
M(\hat\beta_0) = \mathbb{E}\|L_A - \hat L_A(\hat\beta_0)\|^2 = \mathbb{E}\|A(\beta-\hat\beta_0)+Z\|^2 &= \operatorname{tr}\left(\operatorname{Cov}\left[A\hat\beta_0 - Z\right]\right) \\
&= \operatorname{tr}\left(\operatorname{Cov}\left[A\hat\beta_0\right]\right) - 2\operatorname{tr}\left(\operatorname{Cov}\left[A\hat\beta_0, Z\right]\right) + \operatorname{tr}\left(\operatorname{Cov}\left[Z,Z\right]\right) \\
&= \operatorname{tr}\left(\operatorname{Cov}\left[\hat\beta_0\right]A^\top A\right) - 2\operatorname{tr}\left(A\operatorname{Cov}\left[\hat\beta_0, Z\right]\right) + hd\sigma^2
\end{aligned}$$

For OLS, we can calculate, recalling $N = \text{Cov}\left[\varepsilon, Z\right]$, $\text{Cov}\left[\hat{\beta}, Z\right] = (X^\top X)^{-1} X^\top N$. Hence

$$M(\hat{\beta}) = \sigma^2 \cdot \left[\text{tr}\left[(X^\top X)^{-1} A^\top A\right] - 2\,\text{tr}\left[A(X^\top X)^{-1} X^\top N\right] + hd\right].$$

For the distributed estimator $\hat{\beta}_{dist}(w) = \sum_i w_i \hat{\beta}_i$, we have

$$\begin{aligned}\text{Cov}\left[\hat{\beta}_{dist}, Z\right] &= \text{Cov}\left[\sum_i w_i (X_i^\top X_i)^{-1} X_i^\top \varepsilon_i, Z\right] \\ &= \sum_i w_i (X_i^\top X_i)^{-1} X_i^\top \,\text{Cov}\left[\varepsilon_i, Z\right] = \sum_i w_i (X_i^\top X_i)^{-1} X_i^\top N_i\end{aligned}$$

Above, we denoted $N_i := \text{Cov}\left[\varepsilon_i, Z\right]$. Therefore,

$$M(\hat{\beta}_{dist}) = \sigma^2 \cdot \left(\sum_{i=1}^k w_i^2 \cdot \text{tr}\left[(X_i^\top X_i)^{-1} A^\top A\right] - 2 w_i \cdot \text{tr}\left[A(X_i^\top X_i)^{-1} X_i^\top N_i\right]\right) + \sigma^2 hd.$$

To find the optimal weights, we consider more generally the quadratic optimization problem

$$\min_{w \in \mathbb{R}^k} \sum_{i=1}^k \frac{a_i}{2} w_i^2 - b_i w_i$$

subject to $\sum_{i=1}^k w_i = 1$. We assume that $a_i > 0$. In that case, the problem is convex, and we can use a simple Lagrangian reformulation to solve it. Note that we do not impose the constraint $w_i \geq 0$, because in principle one could allow negative weights, and because it is usually satisfied without imposing the constraint.

Denoting by $\Psi(w)$ the objective, we consider the problem of minimizing the Lagrangian $\Psi_\lambda(w) = \Psi(w) - \lambda(\sum_i w_i - 1)$. It is easy to check that the condition $\frac{\partial \Psi_\lambda}{\partial w_i} = 0$ reduces to

$$w_i = \frac{\lambda + b_i}{a_i}.$$

In order for the constraint $\sum_{i=1}^{k} w_i = 1$ to be satisfied, we need that

$$\lambda = \lambda^* := \frac{1 - \sum_{i=1}^{k} \frac{b_i}{a_i}}{\sum_{i=1}^{k} \frac{1}{a_i}}.$$

Plugging back this value of $\lambda$, we obtain the optimal value or the weights $w_i^*$. To apply this result to our problem, we choose $a_i = \operatorname{tr}\left[(X_i^\top X_i)^{-1} A^\top A\right]$, and $b_i = \operatorname{tr}\left(A(X_i^\top X_i)^{-1} X_i^\top N_i\right)$. This finishes the proof.

*2.8.4. Proof of Theorem 2.4.1*

We want to show

$$\widehat{\Sigma}^{-1} \asymp \Sigma^{-1} \cdot e_p.$$

As mentioned, the proof of this result relies on the generalized Marchenko-Pastur theorem of Rubio and Mestre (2011). From that result, we have under the stated assumptions

$$(\widehat{\Sigma} - zI)^{-1} \asymp (x_p \Sigma - zI)^{-1},$$

where $x_p = x_p(z), e_p = e_p(z)$ are the unique solutions of the system

$$e_p = \frac{1}{p} \operatorname{tr}\left[\Sigma(x_p \Sigma - zI)^{-1}\right], \; x_p = \frac{1}{n} \operatorname{tr}\left[\Gamma(I + \gamma_p e_p \Gamma)^{-1}\right].$$

From section 2.2 of Paul and Silverstein (2009), when the e.s.d of $\Sigma$ converges to $H$ and the e.s.d of $\Gamma$ converges to $G$, $x_p$ and $e_p$ will converge to $x$ and $e$ respectively, where $x = x(z)$ and $e = e(z)$ are the unique solutions of the system

$$e = \int \frac{t}{tx - z} dH(t), \; x = \int \frac{t}{1 + \gamma t e} dG(t).$$

Recall that, in Section 2.8.8, we have the system of equations

$$\delta = \int \frac{\gamma t}{-z(1+\tilde{\delta}t)} d\mu_H(t),\ \tilde{\delta} = \int \frac{t}{-z(1+\delta t)} d\mu_G(t).$$

Then, it's easy to check that $\delta = \gamma e$ and $x = -z\tilde{\delta}$. We will use these relations later.

Now, we want to show that we can take $z = 0$, i.e.

$$\widehat{\Sigma}^{-1} \asymp (x_p(0)\Sigma)^{-1} = \Sigma^{-1} \cdot e_p(0).$$

So for a given sequence of matrices $C_p$ with bounded trace norm we need to bound

$$\begin{aligned}
\Delta_p :&= \mathrm{tr}[C_p(\widehat{\Sigma}^{-1} - (x_p(0)\Sigma)^{-1})] \\
&= \mathrm{tr}[C_p(\widehat{\Sigma}^{-1} - (\widehat{\Sigma} - zI)^{-1})] + \mathrm{tr}[C_p((\widehat{\Sigma} - zI)^{-1} - (x_p\Sigma - zI)^{-1}))] \\
&+ \mathrm{tr}[C_p((x_p\Sigma - zI)^{-1} - (x_p\Sigma)^{-1})] + \mathrm{tr}[C_p((x_p\Sigma)^{-1} - (x_p(0)\Sigma)^{-1})] \\
&= \Delta_p^1 + \Delta_p^2 + \Delta_p^3 + \Delta_p^4.
\end{aligned}$$

We can bound the four error terms in turn:

1. Bounding $\Delta_p^1$:

   We have

   $$D_1(z) = (\widehat{\Sigma} - zI)^{-1} - \widehat{\Sigma}^{-1} = z(\widehat{\Sigma} - zI)^{-1}\widehat{\Sigma}^{-1}.$$

   Hence, the operator norm of $D(z)$ can be bounded as

   $$\|D_1(z)\|_{op} \leq \frac{2|z|}{\lambda_{\min}(\widehat{\Sigma})^2}$$

   for sufficiently small $z$.

Recall that $X = \Gamma^{1/2} Z \Sigma^{1/2}$, where $\Gamma$ is a diagonal matrix with positive entries and $\Sigma$ is a symmetric positive definite matrix. Let us consider the least singular value of $X$. By assumption, the entries of $\Gamma$ and the eigenvalues of $\Sigma$ are uniformly bounded below by some constant K. So we can bound $\sigma_{\min}(X)$ as follows:

$$\sigma_{\min}(X) = \sigma_{\min}(\Gamma^{1/2} Z \Sigma^{1/2}) \geq \sigma_{\min}(\Gamma^{1/2}) \sigma_{\min}(Z) \sigma_{\min}(\Sigma^{1/2}) \geq K \cdot \sigma_{\min}(Z).$$

By using the bound above, we have

$$\begin{aligned} \lambda_{\min}(\hat{\Sigma}) = \lambda_{\min}(\frac{X^\top X}{n}) &= \frac{(\sigma_{\min}(X))^2}{n} \geq \frac{K^2 \cdot (\sigma_{\min}(Z))^2}{n} \\ &= K^2 \cdot \lambda_{\min}(\frac{Z^\top Z}{n}) \to_{a.s.} K^2 (1 - \sqrt{\gamma})^2, \end{aligned}$$

where the final step comes from the well-known Bai-Yin law (Bai and Silverstein, 2010).

Thus, we conclude that

$$\begin{aligned} \lim_{p \to +\infty} |\Delta_p^1| &= \lim_{p \to +\infty} |\operatorname{tr}[C_p(\widehat{\Sigma}^{-1} - (\widehat{\Sigma} - zI)^{-1})]| \\ &\leq \lim_{p \to +\infty} \|C_p\|_{tr} \cdot \|D_1(z)\|_{op} \leq \lim_{p \to +\infty} \|C_p\|_{tr} \cdot \frac{2|z|}{(K^2 \cdot \lambda_{\min}(Z^\top Z/n))^2} \leq C'|z|. \end{aligned}$$

This holds almost surely, for some fixed constant $C' > 0$.

2. Bounding $\Delta_p^2$:

   This just follows Theorem 1 of Rubio and Mestre (2011):

$$|\Delta_p^2| = \operatorname{tr}[C_p((\widehat{\Sigma} - zI)^{-1} - (x_p \Sigma - zI)^{-1}))] \to_{a.s.} 0.$$

3. Bounding $\Delta_p^3$:

   By a similar logic, we can obtain a bound on the operator norm of

$$D_2(z) = (x_p\Sigma - zI)^{-1} - (x_p\Sigma)^{-1}$$

for sufficiently small $z$, of the form

$$\|D_2(z)\|_{op} \leq \frac{2|z|}{|x_p(z)|^2 \cdot \lambda_{\min}(\Sigma)^2}$$

for sufficiently small $z$. Again, we have assumed that the smallest eigenvalues of $\Sigma$ are always bounded away from zero, so that $\lambda_{\min}(\Sigma) > c > 0$ for some fixed constant $c > 0$. Since $x_p(z) \to x(z) = -z\tilde{\delta}(z)$ as $p \to +\infty$, and we know that $-z\tilde{\delta}(z)$ is analytic in a neighborhood of the origin with $x(0) = \lim_{z\to 0}[-z\tilde{\delta}(z)] = \tilde{\rho}(\{0\}) > 0$. We can argue that $|x(z)|$ is bounded below in a neighborhood of the origin.

So we conclude that

$$\begin{aligned}
\lim_{p\to+\infty} |\Delta_p^3| &= \lim_{p\to+\infty} |\operatorname{tr}[C_p((x_p\Sigma - zI)^{-1} - (x_p\Sigma)^{-1})]| \\
&\leq \lim_{p\to+\infty} \|C_p\|_{tr} \cdot \|D_2(z)\|_{op} \\
&\leq \limsup \|C_p\|_{tr} \cdot \frac{2|z|}{|x(z)|^2 \cdot \lambda_{\min}(\Sigma)^2} \leq C''|z|.
\end{aligned}$$

This holds almost surely, for some fixed constant $C'' > 0$.

4. Bounding $\Delta_p^4$:

$$\begin{aligned}
\lim_{p\to+\infty} |\Delta_p^4| &= \lim_{p\to+\infty} |\operatorname{tr}[C_p((x_p\Sigma)^{-1} - (x_p(0)\Sigma)^{-1})] \\
&\leq \lim_{p\to+\infty} \|C_p\|_{tr} \cdot \frac{|x_p(z)^{-1} - x_p(0)^{-1}|}{\lambda_{\min}(\Sigma)} \\
&\leq \limsup \|C_p\|_{tr} \cdot \frac{|x(z)^{-1} - x(0)^{-1}|}{\lambda_{\min}(\Sigma)} \leq C'''|z|.
\end{aligned}$$

This holds almost surely for some fixed constant $C''' > 0$, since $x(z)$ is analytic near the origin with $x(0) > 0$.

Combining these, we have

$$\lim_{p\to+\infty} |\Delta_p| = \lim_{p\to+\infty} |\Delta_p^1 + \Delta_p^2 + \Delta_p^3 + \Delta_p^4| \leq (C' + C'' + C''')|z|.$$

Since $|z|$ can be arbitrarily small, we conclude that, almost surely

$$\lim_{p\to+\infty} |\Delta_p| = \lim_{p\to+\infty} \mathrm{tr}[C_p(\widehat{\Sigma}^{-1} - (x_p(0)\Sigma)^{-1})] = 0.$$

This finishes the proof.

*2.8.5. Proof of Theorem 2.4.3*

Recall that we defined $A_n \asymp B_n$ if $\lim_{n\to\infty} |\mathrm{tr}\,[E_n(A_n - B_n)]| = 0$ a.s., for any standard sequence $E_n$ (of symmetric deterministic matrices with bounded trace norm). Below, $E_n$ will always denote such a sequence.

1. The three required properties are that the $\asymp$ relation is reflexive, symmetric and transitive. The reflexivity and symmetry are obvious. To verify transitivity, we suppose $A_n \asymp B_n$ and $B_n \asymp C_n$. Then, for any standard sequence $E_n$, by the triangle inequality,

$$|\mathrm{tr}\,[E_n(A_n - C_n)]| \leq |\mathrm{tr}\,[E_n(A_n - B_n)]| + |\mathrm{tr}\,[E_n(B_n - C_n)]|\,.$$

   Since the two sequences on the right hand side converge to zero almost surely, the conclusion follows.

2. Let $D_n^1 = A_n - B_n$ and $D_n^2 = C_n - D_n$. Then we can bound by the triangle inequality

$$\left|\mathrm{tr}\left[E_n(D_n^1 + D_n^2)\right]\right| \leq \left|\mathrm{tr}\left[E_n D_n^1\right]\right| + \left|\mathrm{tr}\left[E_n D_n^2\right]\right|.$$

As before, the two sequences on the right hand side converge to zero almost surely, so the conclusion follows.

3. We need to show that $A_nB_n \asymp A_nC_n$. Let $E_n$ be any standard sequence. For this it is enough to show that $A_nE_n$ is still a standard sequence. However, this is clear, because

$$\limsup \|A_nE_n\|_{tr} \leq \limsup \|A_n\|_{op}\|E_n\|_{tr} \leq \limsup \|A_n\|_{op} \limsup \|E_n\|_{tr} < \infty.$$

4. We know that $\lim_{n\to\infty} |\mathrm{tr}\,[E_n(A_n - B_n)]| = 0$ a.s., for any standard sequence $E_n$. Consider $E_n = n^{-1}I_n$. Then $\|E_n\|_{tr} = 1$, so that $E_n$ is a standard sequence. Therefore, $\lim_{n\to\infty} |\mathrm{tr}\,[A_n - B_n]| = 0$ a.s., as desired.

5. This is a direct consequence of the trace property.

*2.8.6. Applications of the calculus*

In this section we briefly sketch several applications of the calculus of deterministic equivalents. We emphasize that in each case, there are other proof techniques, but they tend to be more case-by-case. The calculus provides a unified set of methods using which separate results can be seen as applications of the same approach.

**Risk of ridge regression.** We can get a simpler derivation of certain previously found formulas for the risk of ridge regression (Dobriban and Wager, 2018). Considering Theorem 2.1 in that work, the finite-sample predictive risk of ridge regression involves $\mathrm{tr}(\Sigma(\widehat{\Sigma}+\lambda I)^{-1})$. Using the calculus of deterministic equivalents, we can compute this quantity using the following steps, starting with the main equivalence:

$$(\widehat{\Sigma} + \lambda I)^{-1} \asymp (x_p\Sigma + \lambda I)^{-1}$$

$$\Sigma(\widehat{\Sigma} + \lambda I)^{-1} \asymp \Sigma(x_p\Sigma + \lambda I)^{-1}$$

$$p^{-1}\,\mathrm{tr}[\Sigma(\widehat{\Sigma} + \lambda I)^{-1}] - p^{-1}\,\mathrm{tr}[\Sigma(x_p\Sigma + \lambda I)^{-1}] \to_{a.s.} 0.$$

It remains to find the limit of the right hand side. This can be done by using the fixed point equation defining $x_p$. From the proof of Theorem 2.4.1, we have that $x_p$ and the associated scalar $e_p$ are the unique solutions of the system

$$e_p = \frac{1}{p} \operatorname{tr}\left[\Sigma(x_p\Sigma + \lambda I)^{-1}\right], \; x_p = (1 + \gamma_p e_p)^{-1}.$$

This gives a characterization of $x_p$ that cannot be simplified, and hence solves the problem to the extent that random matrix theory can, recovering the known results in a simpler way (Dobriban and Wager, 2018).

**Fine-grained structure of ridge regression.** In a recent work (Liu and Dobriban, 2019), we have studied ridge regression on a deeper level, including presenting an equivalent for ridge as a sum of a covariance matrix dependent transform of the parameter vector, and another transform of the noise. In that work, we have relied significantly on the calculus of deterministic equivalents.

**Distributed ridge regression.** In a follow-up work to the present one, we study one-shot distributed ridge regression (Dobriban and Sheng, 2019). In that work, we rely heavily on the calculus of deterministic equivalents, and we think that this is a good example of results that would be hard or complicated to obtain otherwise. We refer the reader to Dobriban and Sheng (2019) for details.

**Gradient flow for least squares.** To study gradient flow for least squares regression, Ali et al. (2019) also require the limiting behavior of $\operatorname{tr}(\Sigma(\widehat{\Sigma} + \lambda I)^{-1})$ (see the proof of their Theorem 6). Our calculus can thus be used as an alternative way to derive their results.

**Interpolation.** The recent work of Hastie et al. (2019) has studied high-dimensional interpolation using techniques from random matrix theory. Some of their arguments can be phrased and simplified in the language of the calculus of deterministic equivalents. Consider for instance their Lemma 2, whose proof in version 3 of their arXiv preprint relies on the results of Rubio and Mestre (2011). This proof can be simplified if expressed in the natural

way in the calculus: In equation 30, they want to find the limit

$$\lim_{z \to 0^+} z\beta^\top (\widehat{\Sigma} + zI)^{-1}\beta.$$

Using the calculus of deterministic equivalents, we know that for any fixed $z$, and any fixed $\beta$-sequence with bounded norm

$$\begin{aligned} z(\widehat{\Sigma} + zI)^{-1} &\asymp z(x_p\Sigma + zI)^{-1} \\ z\beta\beta^\top(\widehat{\Sigma} + zI)^{-1} &\asymp z\beta\beta^\top(x_p\Sigma + zI)^{-1} \\ z\beta^\top(\widehat{\Sigma} + zI)^{-1}\beta - z\beta^\top(x_p\Sigma + zI)^{-1}\beta &\to_{a.s.} 0. \end{aligned}$$

These results are not uniform in $z$, but this could be proved with a bit more effort. Now, in their work $\Sigma = I_p$, hence

$$z\beta^\top(x_p\Sigma + zI)^{-1}\beta = z/(x_p + z)\|\beta\|^2.$$

Moreover,

$$e_p = 1/(x_p + z),\ x_p = 1/(1 + \gamma_p e_p).$$

After some elementary calculation, we find that the limit as $z \to 0^+$ when $\|\beta\|^2 = r^2$ is $(1 - 1/\gamma)r^2$, which agrees with Hastie et al. (2019).

**Heteroskedastic PCA.** The calculus of deterministic equivalents can be used to simplify certain arguments used to study heteroskedastic PCA (Hong et al., 2018a,b). Specifically, for Lemma 5 in Hong et al. (2018b), the key problem is to find the limit of $\zeta\,\mathrm{tr}[WRW]$, where $R = (\zeta^2 I - \tilde{E}^H\tilde{E})^{-1}$ is the resolvent of a Marchenko-Pastur type matrix $\tilde{E}^H\tilde{E}$. The calculus of deterministic equivalents leads, using notation that has to be changed mutatis

mutandis, and $\Sigma := \mathbb{E}\tilde{E}^H\tilde{E}$, to

$$\zeta(\zeta^2 I - \tilde{E}^H\tilde{E})^{-1} \asymp \zeta(\zeta^2 I - x_p\Sigma)^{-1}$$
$$\zeta W(\zeta^2 I - \tilde{E}^H\tilde{E})^{-1}W \asymp \zeta W(\zeta^2 I - x_p\Sigma)^{-1}W$$
$$n^{-1}\zeta \operatorname{tr} W(\zeta^2 I - \tilde{E}^H\tilde{E})^{-1}W - n^{-1}\zeta \operatorname{tr} W(\zeta^2 I - x_p\Sigma)^{-1}W \to_{a.s.} 0.$$

Then, using the specific expression of $\Sigma$, which is diagonal in this case, it is not hard to recover the statement of Lemma 5 from Hong et al. (2018b).

**ePCA theory.** Another application of the calculus of deterministic equivalents is to develop a rigorous analysis for spiked covariance models in exponential families, which were proposed in Liu et al. (2018a). This is an ongoing project of one of the authors, and we refer to the forthcoming manuscript for details (Dobriban et al., 2019).

*2.8.7. Elliptical models*

We study the setting of elliptical data. In this model the data samples may have different scalings, having the form $x_i = g_i^{1/2}\Sigma^{1/2}z_i$, for some vector $z_i$ with iid entries, and for datapoint-specific *scale parameters* $g_i$. Arranging the data as the rows of the matrix $X$, that takes the form

$$X = \Gamma^{1/2}Z\Sigma^{1/2},$$

where $Z$ and $\Gamma$ are as before: $Z$ has iid standardized entries, while $\Sigma$ is the covariance matrix of the features. Now $\Gamma$ is the diagonal *scaling matrix* containing the scales $g_i$ of the samples. This model has a long history in multivariate statistical analysis (e.g., Mardia et al., 1979).

In the elliptical model, we find the following expression for the ARE.

**Theorem 2.8.1** (ARE for elliptical models)**.** *Consider the above high-dimensional asymptotic limit, where the data matrix is random, and the samples have the form $X = \Gamma^{1/2}Z\Sigma^{1/2}$. Suppose that, as $n_i \to \infty$ with $p/n_i \to \gamma_i \in (0,1)$, the e.s.d. of $\Gamma$ converges weakly to $G$,*

*the e.s.d. of each $\Gamma_i$ converges weakly to some $G_i$, and that the e.s.d. of $\Sigma$ converges weakly to $H$. Suppose that $H$ is compactly supported away from the origin, $G$ is also compactly supported and does not have point mass at the origin. Then, the ARE has the form*

$$ARE = f(\gamma, G) \cdot \sum_{i=1}^{k} \frac{1}{f(\gamma_i, G_i)}.$$

See the following sections (Section 2.8.1) for the proof.

There are two implicit relations in the above formula. First, $\sum 1/\gamma_i = 1/\gamma$, because $\sum n_i/p = n/p$. Second, $n \cdot G = \sum_{i=1}^{k} n_i \cdot G_i$, or equivalently $G/\gamma = \sum_{i=1}^{k} G_i/\gamma_i$, because $\Gamma$ contains all entries of each $\Gamma_i$.

For the special case when all aspect ratios $\gamma_i$ are equal, and all scale distributions $G_i$ are equal to $G$, we can say more about the ARE. We have the following theorem.

**Theorem 2.8.2** (Properties of ARE for elliptical models)**.** *Consider the behavior of distributed regression in elliptical models under the conditions of Theorem 2.8.1. Suppose that the data sizes $n_i$ on all machines are equal, so that $\gamma_i = \gamma_j = k\gamma$ for all $i, j$. Suppose moreover that the scale distributions $G_i$ on all machines are also equal. Then, the ARE has the following properties*

*1. It can be expressed equivalently as*

$$ARE(k) = \frac{k \cdot \eta_G^{-1}(1-\gamma)}{\eta_G^{-1}(1-k\gamma)} = \frac{k \cdot f(\gamma, G)}{f(k\gamma, G)} = \frac{e(\gamma, G)}{e(k\gamma, G)}.$$

*Here $\eta_G$ is the $\eta$-transform of $G$, $f$ is defined above, while $e$ is the unique positive solution of the equation*

$$\int \frac{se}{1+\gamma se} dG(s) = 1.$$

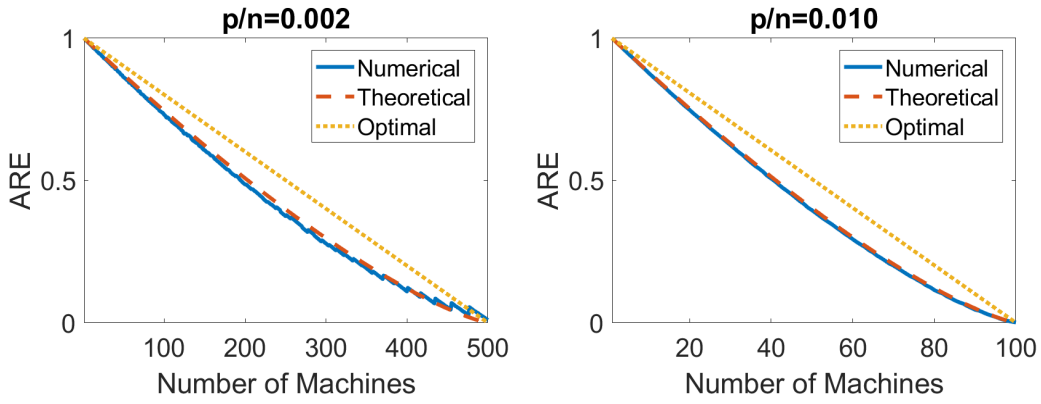



Figure 6: Comparison of empirical and theoretical ARE for elliptical distributions. Left: $n = 10,000$, $p = 20$. Right: $n = 10,000$, $p = 100$.

2. *Suppose also that $G$ does not have a point mass at the origin. Then, the ARE is a strictly decreasing smooth convex function for $k \in [1, 1/\gamma]$. Here $k$ is viewed as a continuous variable. Moreover $ARE(1) = 1$, and*

$$\lim_{k \to 1/\gamma} ARE(k) = 0.$$

See Section 2.8.9 for the proof. These theoretical formulas again match simulation results well, see Figure 6. On that figure, we use the same simulation setup as for Figure 3, and in addition we choose the scale distribution to be uniform on $[0, 1]$.

The ARE for a constant scale distribution is a straight line in $k$, $\text{ARE}(k) = (1-k\gamma)/(1-\gamma)$. For a general scale distribution, the graph of ARE is a curve below that straight line. The interpretation is that for elliptical distributions, there is a larger efficiency loss in one-step averaging. Intuitively, the problem becomes "more non-orthogonal" due to the additional variability from sample to sample.

*2.8.8. Proof of Theorem 2.8.1*

Consider the matrix

$$\widehat{\Sigma} = \frac{1}{n} X^\top X = \frac{1}{n} \Sigma^{\frac{1}{2}} Z^\top \Gamma Z \Sigma^{\frac{1}{2}}.$$

Recall that the *e.s.d.* of $\Gamma$ converges to $G$, and that the *e.s.d.* of $\Sigma$ converges to $H$. According to Paul and Silverstein (2009), with probability 1, the *e.s.d.* of $\widehat{\Sigma}$ converges to a distribution $F$, whose Stieltjes transform $m(z), z \in \mathbb{C}^+$ is given by

$$m(z) = \int \frac{1}{t \int \frac{s}{1+\gamma s e} dG(s) - z} dH(t),$$

where $e = e(z)$ is the unique solution in $\mathbb{C}^+$ of the equation

$$e = \int \frac{t}{t \int \frac{s}{1+\gamma s e} dG(s) - z} dH(t).$$

Since $\mathrm{tr}[(X^\top X)^{-1}] \to \gamma m(0)$, we only need to solve for $m(0)$. As we will show below, we can take $z \to 0$, and obtain that

$$m(0) = \frac{\int \frac{1}{t} dH(t)}{\int \frac{s}{1+\gamma s e} dG(s)}, \quad e(0) = \frac{1}{\int \frac{s}{1+\gamma s e} dG(s)},$$

hence $e := e(0)$ can be checked to be the unique positive solution to the equation

$$\int \frac{se}{1+\gamma s e} dG(s) = 1.$$

To make this rigorous, we need to use some results from Couillet and Hachem (2014). Let $\mu_F, \mu_G, \mu_H$ be the probability measures corresponding to the distributions $F, G, H$. Our goal is to show that, when $\mu_H$ is compactly supported away from the origin, $\mu_G$ is compactly supported and does not have a point mass at the origin, then $\mu_F$ is also compactly supported away from the origin and the solutions $m(z)$, $e(z)$ to the above equations can be extended to the origin.

First, for any $z \in \mathbb{C}^+$, Couillet and Hachem (2014) showed that the system of equations

$$\delta = \int \frac{\gamma t}{-z(1+\tilde{\delta} t)} d\mu_H(t), \ \tilde{\delta} = \int \frac{t}{-z(1+\delta t)} d\mu_G(t)$$

admits a unique solution $(\delta, \tilde{\delta}) \in (\mathbb{C}^+)^2$. Let $\delta(z)$ and $\tilde{\delta}(z)$ be these solutions. Notice that

$\delta(z)$ and $e(z)$ have the following relation: $\delta(z) = \gamma e(z)$. Therefore, we can equivalently study $\delta(z)$ instead of $e(z)$. The function $m(z)$, which is the Stieltjes transform of $\mu_F$, can be expressed as:

$$m(z) = \int \frac{1}{-z(1+\tilde{\delta}(z)t)} d\mu_H(t), \quad z \in \mathbb{C}^+.$$

We will use this expression later.

A important and useful proposition from Couillet and Hachem (2014) is that the functions $\delta(z), \tilde{\delta}(z)$ admit the representations

$$\delta(z) = \int_0^\infty \frac{1}{t-z} d\rho(t), \quad \tilde{\delta}(z) = \int_0^\infty \frac{1}{t-z} d\tilde{\rho}(t),$$

where $\rho$ and $\tilde{\rho}$ are two Radon positive measures on $\mathbb{R}^+$ such that

$$0 < \int_0^\infty \frac{1}{1+t} d\rho(t) < \infty, \quad 0 < \int_0^\infty \frac{1}{1+t} d\tilde{\rho}(t) < \infty.$$

Thus, $\delta(z)$ and $\tilde{\delta}(z)$ can be analytically extended to $\mathbb{C}\backslash\text{supp}(\rho)$ and $\mathbb{C}\backslash\text{supp}(\tilde{\rho})$ respectively.

For the support of measures $\mu_F, \mu_G, \mu_H, \rho$ and $\tilde{\rho}$, we have the following relations from Couillet and Hachem (2014):

1.
$$\mu_F(\{0\}) = 1 - \min[1 - \mu_H(\{0\}), \frac{1-\mu_G(\{0\})}{\gamma}].$$

   So under our assumption that each of $H, G$ have zero point mass at the origin, and that $\gamma < 1$, we have $\mu_F(\{0\}) = 0$.

2. Let $\mathbb{R}^* = \mathbb{R} \setminus \{0\}$, then $\text{supp}(\rho) \cap \mathbb{R}^* = \text{supp}(\tilde{\rho}) \cap \mathbb{R}^* = \text{supp}(\mu_F) \cap \mathbb{R}^*$.

3. Suppose $\inf(\text{supp}(\mu_H) \cap \mathbb{R}^*) > 0$, i.e. the support of $\mu_H$ is away from the origin, then $\inf(\text{supp}(\mu_F) \cap \mathbb{R}^*) > 0$, the support of $\mu_F$ is also away from the origin.

4. $\text{supp}(\mu_F)$ is compact if and only if $\text{supp}(\mu_G)$ and $\text{supp}(\mu_H)$ are both compact.

5. Under our assumption, $\tilde{\rho}(\{0\}) = \lim_{y\downarrow 0}(-iy\tilde{\delta}(iy)) > 0$. Since $\tilde{\delta}(z) \to \infty$ as $z \to 0$ and $\mu_H$ is compactly supported away from the origin, by the dominated convergence theorem (DCT),

$$\rho(\{0\}) = \lim_{y\downarrow 0}(-iy\delta(iy)) = \lim_{y\downarrow 0}\int \frac{\gamma t}{1+\tilde{\delta}(iy)t} d\mu_H(t) = 0.$$

Given these, the picture is now clear. That is, under our assumption, $\text{supp}(\mu_F) = \text{supp}(\rho) = K$, $\text{supp}(\tilde{\rho}) = \{0\} \cup K$, where $K$ is some compact set on $\mathbb{R}^+$ away from the origin. Thus, $m(z)$ and $\delta(z)$ can be analytically extended to $\mathbb{C} \setminus K$. And $\tilde{\delta}(z)$ can be extended to a meromorphic function on $\mathbb{C} \setminus K$, with a simple pole at $z = 0$.

Let us rewrite the system of equations as

$$\delta(z) = \int \frac{\gamma t}{-z(1+\tilde{\delta}(z)t)} d\mu_H(t), \; -z\tilde{\delta}(z) = \int \frac{t}{1+\delta(z)t} d\mu_G(t),$$

where $z \in \mathbb{C}^+$. Now, using the integral representations of $\delta, \tilde{\delta}$ given above,

$$\delta(0) = \int_0^\infty \frac{1}{t} d\rho(t) > 0, \quad \lim_{z\to 0} z\tilde{\delta}(z) = -\tilde{\rho}(\{0\}) < 0,$$

it is easy to check that the right-hand sides of both equations above are analytic at least in a small neighborhood $U$ of the origin. By the uniqueness property of analytic functions, the above system of equations will hold for all $z \in U$. This means that we can evaluate the equation at $z = 0$. For the equation

$$m(z) = \int \frac{1}{-z(1+\tilde{\delta}(z)t)} d\mu_H(t), \quad z \in \mathbb{C}^+,$$

we find by a similar argument that we can also evaluate $m(z)$ at $z = 0$. This finishes the proof for the expressions of $m(0), e(0)$ given at the beginning of the proof of the main theorem.

Moreover, the Stieltjes transform we are looking for has the form $m(0) = e(0) \cdot \mathbb{E}_H T^{-1}$. Let us write $e(\gamma, G)$ for $e(0)$ showing the dependence on $\gamma, G$ explicitly. Then,

$$\operatorname{tr}\left[(X^\top X)^{-1}\right] \to \gamma e(\gamma, G) \cdot \mathbb{E}_H T^{-1}.$$

Similarly, since $X_i$ has the same elliptical form $X_i = \Gamma_i^{1/2} Z_i \Sigma^{1/2}$, and by assumption the e.s.d. of $\Gamma_i$ converges to $G_i$, we obtain that

$$\operatorname{tr}\left[(X_i^\top X_i)^{-1}\right] \to \gamma_i e(\gamma_i, G_i) \cdot \mathbb{E}_H T^{-1}.$$

Thus, the ARE equals

$$ARE = \gamma e(\gamma, G) \cdot \mathbb{E}_H T^{-1} \cdot \sum_{i=1}^{k} \frac{1}{\gamma_i e(\gamma_i, G_i) \cdot \mathbb{E}_H T^{-1}} = \gamma e(\gamma, G) \cdot \sum_{i=1}^{k} \frac{1}{\gamma_i e(\gamma_i, G_i)}.$$

Notice now that $f(\gamma, G) = \gamma e(\gamma, G)$ so the ARE also equals $f(\gamma, G) \cdot \sum_{i=1}^{k} \frac{1}{f(\gamma_i, G_i)}$. This finishes the proof.

*2.8.9. Proof of Theorem 2.8.2*

Under the assumptions of the theorem, we have:

$$ARE(k) = \frac{k f(\gamma, G)}{f(k\gamma, G)} = \frac{k \cdot \eta_G^{-1}(1-\gamma)}{\eta_G^{-1}(1-k\gamma)} = \frac{e(\gamma, G)}{e(k\gamma, G)}.$$

The second form given in the theorem follows directly from the definition of $e$.

Next, we assume $G$ does not have a point mass at the origin. From the definition of $\eta$-transform, we have the following observation. For any $G$, $\eta_G(x)$ is a smooth decreasing function on $[0, +\infty)$ with $\eta_G(0) = 1$ and $\lim_{x \to +\infty} \eta_G(x) = 0$. So $\eta_G^{-1}(x)$ defined on $(0, 1]$ is also smooth and decreasing with $\eta_G^{-1}(1) = 0$ and $\lim_{x \to 0+} \eta_G^{-1}(x) = +\infty$. This means that $ARE(k)$ is indeed a well-defined function for $k \in [1, 1/\gamma]$.

Next we show that ARE($k$) is a decreasing convex function. Convexity is equivalent to saying that $1/e(k\gamma, G)$ is decreasing and convex in $k$. Let $\psi(k) = 1/e(k\gamma, G)$. Then $\psi(k)$ is the unique positive solution to the equation

$$\int \frac{t}{\psi(k) + k\gamma t} dG(t) = 1.$$

We can differentiate with respect to $k$ on both sides to get

$$\psi'(k) = -\frac{\int \frac{\gamma t^2}{(\psi+k\gamma t)^2} dG(t)}{\int \frac{t}{(\psi+k\gamma t)^2} dG(t)} \leq 0.$$

Similarly, we differentiate it twice to get

$$\psi''(k) = \frac{\int \frac{2t(\psi'+\gamma t)^2}{(\psi+k\gamma t)^3} dG(t)}{\int \frac{t}{(\psi+k\gamma t)^2} dG(t)} \geq 0.$$

This proves that $\psi$ is decreasing and convex, and finishes the proof.

*2.8.10. Proof of Theorem 2.5.2*

In the proof of Theorem 2.5.3, we derive the limit for

$$\frac{\mathrm{tr}((X_i^\top X_i)^{-1} X^\top X)}{p} = \frac{p + \sum_{j \neq i} \mathrm{tr}((X_i^\top X_i)^{-1} X_j^\top X_j)}{p},$$

which is

$$1 + (\frac{1}{\gamma} \mathbb{E}_G T - \frac{1}{\gamma_i} \mathbb{E}_{G_i} T) f(\gamma_i, G_i).$$

Then the desired result follows. It is not hard to check the results in the MP case.

*2.8.11. Proof of Theorem 2.5.3*

We consider the Elliptical type sample covariance matrices first. Recall that we have $X = \Gamma^{1/2} Z \Sigma^{1/2}$, where $Z$ is an $n \times p$ matrix with standardized entries, $\Gamma$ is an $n \times n$ diagonal matrix with positive entries and $\Sigma$ is a $p \times p$ nonnegative-definite matrix. Our goal is to

understand the limit of

$$\mathrm{tr}[(X_i^\top X_i)^{-1}X^\top X] = \mathrm{tr}[(X_i^\top X_i)^{-1}X_i^\top X_i]+\sum_{j\neq i}\mathrm{tr}[(X_i^\top X_i)^{-1}X_j^\top X_j] = p+\sum_{j\neq i}\mathrm{tr}[(X_i^\top X_i)^{-1}X_j^\top X_j].$$

If we delete all rows of $X_i$ from $X$ and denote the remaining matrix by $\tilde{X}_i$, then this can be written as

$$p + \mathrm{tr}[\tilde{X}_i(X_i^\top X_i)^{-1}\tilde{X}_i^\top].$$

Since $X_i = \Gamma_i^{1/2}Z_i\Sigma^{1/2}, \tilde{X}_i = \tilde{\Gamma}_i^{1/2}\tilde{Z}_i\Sigma^{1/2}$, where the $n_i \times p$ matrix $Z_i$ and the $(n-n_i)\times p$ matrix $\tilde{Z}_i$ both have i.i.d. standardized entries, the $(n-n_i)\times(n-n_i)$ diagonal matrix $\tilde{\Gamma}_i$ is the remaining matrix after deleting all the entries of $\Gamma_i$ from $\Gamma$. Then we find that the population covariance $\Sigma$ will cancel out:

$$\begin{aligned}&\mathrm{tr}[\tilde{X}_i(X_i^\top X_i)^{-1}\tilde{X}_i^\top] = \mathrm{tr}[\tilde{\Gamma}_i^{1/2}\tilde{Z}_i\Sigma^{1/2}(\Sigma^{1/2}Z_i^\top\Gamma_i Z_i\Sigma^{1/2})^{-1}\Sigma^{1/2}\tilde{Z}_i^\top\tilde{\Gamma}_i^{1/2}]\\ &= \mathrm{tr}[\tilde{\Gamma}_i^{1/2}\tilde{Z}_i(Z_i^\top\Gamma_i Z_i)^{-1}\tilde{Z}_i^\top\tilde{\Gamma}_i^{1/2}].\end{aligned}$$

To evaluate the limit, we will use the following lemma from Rubio and Mestre (2011).

**Lemma 2.8.3** (Concentration of average of quadratic forms, Lemma 4 in Rubio and Mestre (2011))**.** *Let $\mathcal{U} = \{\xi_k \in \mathbb{C}^M, 1 \le k \le N\}$ denote a collection of i.i.d. random vectors with i.i.d entries that have mean 0, variance 1 and finite $4+\delta$ moment, $\delta > 0$. Furthermore, consider a collection of random matrices $\{\mathbf{C}_{(k)} \in \mathbb{C}^{M\times M}, 1 \le k \le N\}$ such that, for each $k$, $\mathbf{C}_{(k)}$ may depend on all the elements of $\mathcal{U}$ except for $\xi_k$, and the trace norm of $\mathbf{C}_{(k)}$, $||\mathbf{C}_{(k)}||_{\mathrm{tr}}$ is almost surely uniformly bounded for all $M$. Then, almost surely as $N \to \infty$,*

$$\left|\frac{1}{N}\sum_{k=1}^{N}\left(\xi_k^H\mathbf{C}_{(k)}\xi_k - \mathrm{tr}\,\mathbf{C}_{(k)}\right)\right| \to 0.$$

For our purpose, we can take the number of summands to be $N = n - n_i$, the dimension

$M = p$, and the inner matrices to be $\mathbf{C}_{(k)} = \frac{n_i}{p}(Z_i^\top \Gamma_i Z_i)^{-1} \cdot (\tilde{\Gamma}_i)_{kk}$. Also, we let $\xi_k^\top$ to be the $k$-th row of $\tilde{Z}_i$. By using the well-known result on spectrum separation(see e.g., Bai and Silverstein, 2010), almost surely, the smallest eigenvalue of ${n_i}^{-1} Z_i^\top Z_i$ is uniformly bounded below by some constant. Since $\lambda_{\min}({n_i}^{-1} Z_i^\top \Gamma_i Z_i) \geq \lambda_{\min}(\Gamma_i) \cdot \lambda_{\min}({n_i}^{-1} Z_i^\top Z_i)$, $\lambda_{\min}({n_i}^{-1} Z_i^\top \Gamma_i Z_i)$ is also uniformly bounded below almost surely. So under the assumption of Theorem 2.5.3, we can check that the trace norm of $\frac{n_i}{p}(Z_i^\top \Gamma_i Z_i)^{-1} \cdot (\tilde{\Gamma}_i)_{kk}$ is indeed uniformly bounded. Then by Lemma 2.8.3, we will have as $n \to \infty$

$$\left| \frac{1}{n - n_i} \sum_{k=1}^{n-n_i} \left[ \frac{n_i}{p} (\tilde{\Gamma}_i)_{kk} \cdot \xi_k^\top (Z_i^\top \Gamma_i Z_i)^{-1} \xi_k - \operatorname{tr}\left( \frac{n_i}{p} (Z_i^\top \Gamma_i Z_i)^{-1} \cdot (\tilde{\Gamma}_i)_{kk} \right) \right] \right| \to_{a.s.} 0.$$

This implies

$$\frac{1}{n - n_i} \operatorname{tr}[\tilde{\Gamma}_i^{1/2} \tilde{Z}_i (Z_i^\top \Gamma_i Z_i)^{-1} \tilde{Z}_i^\top \tilde{\Gamma}_i^{1/2}] \to_{a.s.} f(\gamma_i, G_i) \left( \frac{\gamma_i}{\gamma_i - \gamma} \mathbb{E}_G T - \frac{\gamma}{\gamma_i - \gamma} \mathbb{E}_{G_i} T \right),$$

since $\operatorname{tr}(Z_i^\top \Gamma_i Z_i)^{-1} \to_{a.s.} f(\gamma_i, G_i)$ and

$$\frac{\sum_k (\tilde{\Gamma}_i)_{kk}}{n - n_i} = \frac{n}{n - n_i} \cdot \frac{\operatorname{tr}(\Gamma)}{n} - \frac{n_i}{n - n_i} \cdot \frac{\operatorname{tr}(\Gamma_i)}{n_i} \to_{a.s.} \left( \frac{\gamma_i}{\gamma_i - \gamma} \mathbb{E}_G T - \frac{\gamma}{\gamma_i - \gamma} \mathbb{E}_{G_i} T \right).$$

This holds for all $i$. Thus, for the elliptical model, we have

$$IE(X_1, \ldots, X_k) = \frac{1 - \frac{p}{n}}{1 - \frac{2p}{n} + \frac{1}{\sum_{i=1}^k \frac{n}{\operatorname{tr}[(X_i^\top X_i)^{-1} X^\top X]}}} \to_{a.s.} \frac{1 - \gamma}{1 - 2\gamma + \frac{1}{\sum_{i=1}^k \psi(\gamma_i, G_i)}},$$

where

$$\psi(\gamma_i, G_i) = \frac{1}{\gamma + (\mathbb{E}_G T - \frac{\gamma}{\gamma_i} \mathbb{E}_{G_i} T) f(\gamma_i, G_i)}.$$

Now, for the MP model, we can simply take $\Gamma$ to be identity matrix, then the above result reduces to

$$IE(X_1, \ldots, X_k) = \frac{1 - \frac{p}{n}}{1 - \frac{2p}{n} + \frac{1}{\sum_{i=1}^k \frac{n}{\operatorname{tr}[(X_i^\top X_i)^{-1} X^\top X]}}} \to_{a.s.} \frac{1 - \gamma}{1 - 2\gamma + \frac{\gamma(1-\gamma)}{1 - k\gamma}},$$

which finishes the proof.

*2.8.12. Proof of Theorem 2.5.4*

We first provide the proof for the MP model. Since $\Sigma$ is positive definite, we have

$$x_t^\top (X_i^\top X_i)^{-1} x_t = z_t^\top \Sigma^{1/2} (\Sigma^{1/2} Z_i^\top Z_i \Sigma^{1/2})^{-1} \Sigma^{1/2} z_t = z_t^\top (Z_i^\top Z_i)^{-1} z_t.$$

This cancellation shows that the test error does not depend on the covariance matrix.

For the null case, we will show below that we have, almost surely

$$z_t^\top (Z_i^\top Z_i)^{-1} z_t \to \frac{\gamma_i}{1-\gamma_i}.$$

Hence, we obtain that

$$OE(x_t; X_1, \ldots, X_k) \to_{a.s.} \frac{1 + \frac{\gamma}{1-\gamma}}{1 + \frac{1}{\sum_{i=1}^k \frac{1-\gamma_i}{\gamma_i}}} = \frac{\frac{1}{1-\gamma}}{1 + \frac{1}{\frac{1}{\gamma} - k}}.$$

Under the elliptical model, we have $x_t^\top (X_i^\top X_i)^{-1} x_t = g_t z_t^\top (Z_i^\top \Gamma_i Z_i)^{-1} z_t$. While $\Sigma$ still cancels out, the scale parameters do not cancel out anymore. Therefore, we must take them into account when taking the limits. However, similarly to the proof of Theorem 2.5.2, we find that, almost surely

$$z_t^\top (Z_i^\top \Gamma_i Z_i)^{-1} z_t \to f(\gamma_i, G_i).$$

Putting these together finishes the proof.

To see the reason for the convergence of quadratic forms, we present a slightly different argument. In fact, Theorem 2.5.4 will still hold if we are only given the $4 + c$-th moment of $z_1$ instead of $8 + c$-th one. This follows by the concentration of quadratic forms $x^\top A x - p^{-1} \operatorname{tr} A \to 0$ for matrices $A$ whose spectral distribution converges, and for vectors $x$ with iid entries. Specifically, we will use the following well-known statement about concentration of

quadratic forms. To use this result, we simply choose $x = z_t/\sqrt{p}$, and $A_p = (Z_i^\top \Gamma_i Z_i/p)^{-1}$, and the desired claim follows.

**Lemma 2.8.4** (Concentration of quadratic forms, consequence of Lemma B.26 in Bai and Silverstein (2010))**.** *Let $x \in \mathbb{R}^p$ be a random vector with i.i.d. entries and $\mathbb{E}[x] = 0$, for which $\mathbb{E}\left[(\sqrt{p}x_i)^2\right] = \sigma^2$ and $\sup_i \mathbb{E}\left[(\sqrt{p}x_i)^{4+\eta}\right] < C$ for some $\eta > 0$ and $C < \infty$. Moreover, let $A_p$ be a sequence of random $p \times p$ symmetric matrices independent of $x$, with uniformly bounded eigenvalues. Then the quadratic forms $x^\top A_p x$ concentrate around their means at the following rate*

$$P(|x^\top A_p x - p^{-1}\sigma^2 \operatorname{tr} A_p|^{2+\eta/2} > C) \leq Cp^{-(1+\eta/4)}.$$

To prove this, we will use the following Trace Lemma quoted from Bai and Silverstein (2010), see also Dobriban et al. (2017) for a similar argument.

**Lemma 2.8.5** ([Trace Lemma, Lemma B.26 of Bai and Silverstein (2010))**.** *Let $y$ be a $p$-dimensional random vector of i.i.d. elements with mean 0. Suppose that $\mathbb{E}\left[y_i^2\right] = 1$, and let $A_p$ be a fixed $p \times p$ matrix. Then*

$$\mathbb{E}\left[|y^\top A_p y - \operatorname{tr} A_p|^q\right] \leq C_q \left\{ \left(\mathbb{E}\left[y_1^4\right] \operatorname{tr}[A_p A_p^\top]\right)^{q/2} + \mathbb{E}\left[y_1^{2q}\right] \operatorname{tr}[(A_p A_p^\top)^{q/2}] \right\},$$

*for some constant $C_q$ that only depends on $q$.*

*Proof.* Under the conditions of Lemma 2.8.4, the operator norms $\|A_p\|_2$ are almost surely uniformly bounded by a constant $C$, thus $\operatorname{tr}[(A_p A_p^\top)^{q/2}] \leq pC^q$ and $\operatorname{tr}[A_p A_p^\top] \leq pC^2$. Consider now a random vector $x$ with the properties assumed in the present lemma. For $y = \sqrt{p}x/\sigma$ and $q = 2 + \eta/2$, using that $\mathbb{E}\left[y_i^{2q}\right] \leq C$ and the other the conditions in Lemma

2.8.4, Lemma 2.8.5 thus yields

$$\frac{p^q}{\sigma^{2q}}\mathbb{E}\left[|x^\top A_p x - \frac{\sigma^2}{p}\operatorname{tr} A_p|^q\right] \leq C\left\{\left(pC^2\right)^{q/2} + pC^q\right\},$$

or equivalently $\mathbb{E}\left[|x^\top A_p x - \frac{\sigma^2}{p}\operatorname{tr} A_p|^{2+\eta/2}\right] \leq Cp^{-(1+\eta/4)}$.

By Markov's inequality applied to the $2+\frac{\eta}{2}$-th moment of $\varepsilon_p = x^\top A_p x - \frac{\sigma^2}{p}\operatorname{tr} A_p$, we obtain as required

$$P(|\varepsilon_p|^{2+\eta/2} > C) \leq Cp^{-(1+\eta/4)}.$$

□

*2.8.13. Proof of Theorem 2.5.5*

From Theorem 2.4.1, it follows that the inverse sample covariance matrix $\widehat{\Sigma}$ is equivalent to a scaled version of the population covariance

$$\widehat{\Sigma}^{-1} \asymp \Sigma^{-1} \cdot e_p,$$

for some scalar sequence $e_p > 0$. By taking in Theorem 2.4.1 the matrix $C_p = E_j E_j^\top$, the $p \times p$ matrix with a 1 in the $(j,j)$-th entry, and zeros otherwise, we find that almost surely,

$$[\widehat{\Sigma}^{-1}]_{jj} - [\Sigma^{-1}]_{jj} \cdot e_p \to 0,$$

We can apply this to each sub-matrix $X_i$ to find

$$n_i \cdot [(X_i^\top X_i)^{-1}]_{jj} - [\Sigma^{-1}]_{jj} \cdot e_p(i) \to 0.$$

Here $e_p(i)$ is the solution to the fixed point equation

$$1 = \frac{1}{n_i}\operatorname{tr}\left[e_p(i)\Gamma_i(I_{n_i} + \gamma_{p,i} \cdot e_p(i)\Gamma_i)^{-1}\right].$$

Moreover, $\gamma_{p,i} = p/n_i$ and $\Gamma_i$ is the $n_i \times n_i$ sub-matrix of $\Gamma$ corresponding to the $i$-th machine. It follows that the CE has a deterministic equivalent equal to

$$\frac{[\Sigma^{-1}]_{jj} \cdot e_p}{n} \cdot \sum_{i=1}^{k} \frac{n_i}{[\Sigma^{-1}]_{jj} \cdot e_p(i)} =$$
$$= \frac{p \cdot e_p}{n} \cdot \sum_{i=1}^{k} \frac{n_i}{p \cdot e_p(i)} \to \gamma \cdot e(\gamma, G) \cdot \sum_{i=1}^{k} \frac{1}{\gamma_i \cdot e(\gamma_i, G_i)}.$$

Here $e(\gamma_i, G_i)$ are the quantities encountered before, discussed after Theorem 2.4.1. The convergence follows from the discussion after Theorem 2.4.1. Also, from the definition of $f(\gamma, G)$ it follows that $f(\gamma, G) = \gamma e(\gamma, G)$, so that we get the desired result. This finishes the proof.

*2.8.14. Suboptimal weights*

If we take all weights $w_i$ to be equal, i.e. $w_i = 1/k$, then the MSE is

$$MSE_{subopt} = \frac{\sigma^2}{k^2} \sum_{i=1}^{k} \mathrm{tr}(X_i^\top X_i)^{-1} \to \frac{\sigma^2}{k^2} \sum_{i=1}^{k} \frac{\gamma_i \cdot \mathbb{E}_H T^{-1}}{1 - \gamma_i}.$$

Thus the ARE of the equally weighted estimator becomes (with the notation AE denoting asymptotic MSE)

$$ARE_{subopt} = \frac{AE(\hat{\beta}_{dist}(1/k, \ldots, 1/k))}{AE_{subopt}} = \frac{k^2 \frac{\gamma}{1-\gamma}}{\sum_{i=1}^{k} \frac{\gamma_i}{1-\gamma_i}}.$$

Now, $ARE_{subopt}$ can be viewed as a harmonic mean of the numbers

$$k \frac{\gamma}{1-\gamma} \frac{1-\gamma_i}{\gamma_i},$$

while the optimal ARE is the corresponding arithmetic mean. Therefore, we have $ARE_{subopt} \leq ARE$.

*2.8.15. Properties and interpretation of the relative efficiency.*

Let $f(n, p, k)$ be the relative efficiency for estimation, $(n - kp)/(n - p)$. If $kp > n$, that expression is negative, but in that case it is more proper to define the relative efficiency as 0. So, we consider

$$f(n, p, k) = \max\left(\frac{n - kp}{n - p}, 0\right).$$

This has the following properties. Each of these has a statistical interpretation.

1. **Well-definedness**. $f$ is well-defined for all $n, p, k$ such that $n > p$
2. **Range**. $0 \leq f \leq 1$ for all $n, p, k$. Clearly the efficiency should be between zero and unity.

   Also, $f$ is zero for $k \geq n/p$. In this case, some machine has an OLS estimator that is not well-defined.

   Moreover, $f = 1$ when $k = 1$ or when $p = 0$. When we have one machine, the efficiency is unity by definition. When $p = 0$, the problem is not well-defined, as there are no parameters to estimate.
3. **Monotonicity**.
   (a) $f$ **is monotone decreasing in** $k$**.** This property is easy to interpret. The distributed regression problem gets harder as $k$ increases.
   (b) $f$ **is monotone increasing in** $n$**.** The linear regression problem should get easier as $n$ grows. However, it turns out that more is true. The distributed problem gets relatively easier compared to the "shared" problem.
   (c) $f$ **is monotone decreasing in** $p$**.** Similarly, a typical linear regression problem should get harder as $p$ grows. However, the relative difficulty of solving the distributed problem also gets larger.
4. **Limits and singularity**.
   (a) $n \to \infty$. When $n \to \infty$ with fixed $k, p$, then $f$ tends to unity. When $n \to \infty$, the distributed estimator becomes asymptotically efficient.
   (b) $p = n$. The function is singular when $p = n$, because the OLS estimator itself is

singular when $p = n$.

Note that these properties are not enough to characterize the relative efficiency. In fact, for any monotone increasing transform such that $g(0) = 0$ and $g(1) = 1$, $g(f(n, p, k))$ has the same properties.

*2.8.16. Degrees of freedom interpretation*

Next we give a multi-response regression characterization that heuristically gives an upper bound on the "*degrees of freedom*" for distributed regression. This will be helpful to understand the asymptotic formulas derived above.

We re-parametrize $Y = X\beta+\varepsilon$, treating the samples on each machine as a different outcome. We write the $n \times k$ multi-response outcome matrix $\underline{Y}$, the $n \times pk$ feature matrix $\underline{X}$, and the corresponding noise $\underline{\varepsilon}$ as

$$\underline{Y} = \begin{bmatrix} Y_1 & 0 & \dots & 0 \\ 0 & Y_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & Y_k \end{bmatrix}, \ \underline{X} = \begin{bmatrix} X_1 & 0 & \dots & 0 \\ 0 & X_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & X_k \end{bmatrix}, \ \underline{\varepsilon} = \begin{bmatrix} \varepsilon_1 & 0 & \dots & 0 \\ 0 & \varepsilon_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \varepsilon_k \end{bmatrix}.$$

We also introduce $\underline{\beta}$, the $pk \times k$ parameter matrix, which shares parameters across the $k$ outcomes:

$$\underline{\beta} = \begin{bmatrix} \beta & 0 & \dots & 0 \\ 0 & \beta & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \beta \end{bmatrix} = I_k \otimes \beta$$

Note that $Y = X\beta + \varepsilon$ is equivalent to $\underline{Y} = \underline{X}\underline{\beta} + \underline{\varepsilon}$. The OLS estimator of $\underline{\beta}$ is $\hat{\underline{\beta}} = (\underline{X}^\top \underline{X})^{-1} \underline{X}^\top \underline{Y}$. This can be calculated as

$$\underline{\hat{\beta}} = \begin{bmatrix} \hat{\beta}_1 & 0 & \dots & 0 \\ 0 & \hat{\beta}_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \hat{\beta}_k \end{bmatrix} = \begin{bmatrix} (X_1^\top X_1)^{-1} X_1^\top Y_1 & 0 & \dots & 0 \\ 0 & (X_2^\top X_2)^{-1} X_2^\top Y_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & (X_k^\top X_k)^{-1} X_k^\top Y_k \end{bmatrix}$$

Notice that the estimators of the coefficients of different outcomes are the familiar distributed OLS estimators. Now, we can find a plug-in estimator of $\beta$, based on $\underline{\beta}$. Given the form of $\underline{\beta}$ above, for any vector $w$ such that $\sum_{i=1}^k w_i = 1$, we have that $\beta$ can be expressed in terms of the tensorized parameter $\underline{\beta}$ as a weighted combination

$$\beta = (1_p^\top \otimes I_k)\underline{\beta} w.$$

Therefore, for any unbiased estimator of $\underline{\beta}$, the corresponding weighted combination estimators given below are unbiased for $\beta$:

$$\hat{\beta}(w) = (1_p^\top \otimes I_k)\underline{\hat{\beta}} w.$$

In our case, given the zeros in the estimator, this simply reduces to the weighted sum $\hat{\beta}(w) = \sum_{i=1}^k w_i \hat{\beta}_i$.

This explains how our problem can be understood in the framework of multi-response regression. Also, the number of parameters in that problem is $kp$, so the "degrees of freedom" is $n - kp$. Indeed, the residual effective degrees of freedom $\hat{r} = y - Hy$ is usually defined as $\mathrm{tr}(I - H)$. Let $H_i$ be the hat matrix on the $i$-th machine, so that $H_i = X_i(X_i^\top X_i)^{-1} X_i^\top$. Then it is easy to see that $\mathrm{tr}(I - H_i) = n_i - p$, for all $i$. Since $H_{dist}$ is simply the block diagonal matrix with $H_i$ as blocks, we see that $\mathrm{tr}(I - H_{dist}) = n - pk$, as required.

This provides a simple explanation for why the "effective number of parameters" in a

one-step distributed linear regression problem is upper bounded by $kp$. Equivalently, the residual "degrees of freedom" of a one-step distributed estimation problem is lower bounded by $n - kp$. Note that with more rounds of communication, we could drive the degrees of freedom up, and hence this is just a heuristic bound.

*2.8.17. Proof of Theorem 2.7.2*

First, when $\rho = 0$, the MSE reduces to $\sigma^2/k^2 \cdot \sum_{i=1}^{k} \mathrm{tr}(X_i^\top X_i)^{-1}$ which is exactly the MSE of the one-step naive average estimator $1/k \cdot \sum_{i=1}^{k} (X_i^\top X_i)^{-1} X_i^\top Y_i$. Second, when $\rho \to \infty$,

$$\begin{aligned}
\hat{\beta}_* &= \left( \sum_{i=1}^{k} \left( X_i^\top X_i + n_i \rho I \right)^{-1} X_i^\top X_i \right)^{-1} \cdot \sum_{i=1}^{k} \left( X_i^\top X_i + n_i \rho I \right)^{-1} X_i^\top Y_i \\
&= \left( \sum_{i=1}^{k} \left( \frac{X_i^\top X_i}{\rho} + n_i I \right)^{-1} X_i^\top X_i \right)^{-1} \cdot \sum_{i=1}^{k} \left( \frac{X_i^\top X_i}{\rho} + n_i I \right)^{-1} X_i^\top Y_i \\
&\to \left( \sum_{i=1}^{k} \frac{X_i^\top X_i}{n_i} \right)^{-1} \cdot \sum_{i=1}^{k} \frac{X_i^\top Y_i}{n_i} = (X^\top X)^{-1} X^\top Y,
\end{aligned}$$

which is the OLS estimator for the whole data set and the MSE also converges to the corresponding MSE $\sigma^2 \cdot \mathrm{tr}(X^\top X)^{-1}$. Actually, we can define the MSE as a function of $\rho$ in the following way

$$\begin{aligned}
\psi(\rho) &= \sum_{i=1}^{k} \mathrm{tr} \left[ \left( \sum_{i=1}^{k} \left( X_i^\top X_i + n_i \rho I \right)^{-1} X_i^\top X_i \right)^{-2} \left( X_i^\top X_i + n_i \rho I \right)^{-2} X_i^\top X_i \right] \\
&= \sum_{i=1}^{k} \mathrm{tr} \left[ \left( \sum_{i=1}^{k} \left( \widehat{\Sigma}_i + \rho I \right)^{-1} \widehat{\Sigma}_i \right)^{-2} \left( \widehat{\Sigma}_i + \rho I \right)^{-2} \cdot \frac{\widehat{\Sigma}_i}{n_i} \right],
\end{aligned}$$

where $\widehat{\Sigma}_i = X_i^\top X_i / n_i$. For any fixed $\rho \in [0, \infty)$, we can consider a small perturbation $\varepsilon$ at $\rho$, i.e. we can consider the difference between $\psi(\rho + \varepsilon)$ and $\psi(\rho)$. By using the formula

$$(I + \varepsilon A)^{-1} = \sum_{n=0}^{\infty} (-1)^n A^n \varepsilon^n = I - \varepsilon A + o(\varepsilon),$$

and we define the following quantity

$$\Delta := \sum_{i=1}^{k} \left(\widehat{\Sigma}_i + \rho I\right)^{-1} \widehat{\Sigma}_i.$$

We have the following expansions:

$$\left(\sum_{i=1}^{k} \left(\widehat{\Sigma}_i + \rho I + \varepsilon I\right)^{-1} \widehat{\Sigma}_i\right)^{-1} = \Delta^{-1} + \varepsilon \Delta^{-1} \left(\sum_{i=1}^{k} \left(\widehat{\Sigma}_i + \rho I\right)^{-2} \widehat{\Sigma}_i\right) \Delta^{-1} + o(\varepsilon),$$

$$\left(\widehat{\Sigma}_i + \rho I + \varepsilon I\right)^{-2} \widehat{\Sigma}_i = \left(\widehat{\Sigma}_i + \rho I\right)^{-2} \widehat{\Sigma}_i - 2\varepsilon \left(\widehat{\Sigma}_i + \rho I\right)^{-3} \widehat{\Sigma}_i + o(\varepsilon).$$

By putting these together, we have

$$\begin{aligned}
&\psi(\rho + \varepsilon) - \psi(\rho) = \\
&\varepsilon \cdot \frac{2k}{n} \operatorname{tr}\left[\Delta^{-1} \sum_{i=1}^{k} \left(\widehat{\Sigma}_i + \rho I\right)^{-2} \widehat{\Sigma}_i \cdot \Delta^{-2} \sum_{i=1}^{k} \left(\widehat{\Sigma}_i + \rho I\right)^{-2} \widehat{\Sigma}_i - \Delta^{-2} \sum_{i=1}^{k} \left(\widehat{\Sigma}_i + \rho I\right)^{-3} \widehat{\Sigma}_i\right] \\
&+ o(\varepsilon),
\end{aligned}$$

which means $\psi(\rho)$ is differentiable and the derivative at $\rho$ is

$$\psi'(\rho) = \frac{2k}{n} \operatorname{tr}\left[\Delta^{-1} \sum_{i=1}^{k} \left(\widehat{\Sigma}_i + \rho I\right)^{-2} \widehat{\Sigma}_i \cdot \Delta^{-2} \sum_{i=1}^{k} \left(\widehat{\Sigma}_i + \rho I\right)^{-2} \widehat{\Sigma}_i - \Delta^{-2} \sum_{i=1}^{k} \left(\widehat{\Sigma}_i + \rho I\right)^{-3} \widehat{\Sigma}_i\right].$$

Actually, we can show that $\psi'(\rho) \leq 0$ holds for all $\rho \in [0, +\infty)$. In order to do that, we need to introduce the so-called Schur complement and its applications on positive semi-definite matrices.

**Lemma 2.8.6** (Schur complement condition for positive semi-definiteness)**.** *For any symmetric matrix $M$ of the form*

$$M = \begin{bmatrix} A & B \\ B^\top & C \end{bmatrix},$$

*if $C$ is invertible then the following properties hold:*

*1. $M \succ 0$ iff $C \succ 0$ and $A - BC^{-1}B^\top \succ 0$.*

*2. If $C \succ 0$, then $M \succeq 0$ iff $A - BC^{-1}B^\top \succeq 0$.*

In order to show $\psi'(\rho) \leq 0$, it is sufficient to show

$$\begin{aligned}
&\operatorname{tr}\left[\Delta^{-1}\sum_{i=1}^{k}\left(\widehat{\Sigma}_i+\rho I\right)^{-2}\widehat{\Sigma}_i\cdot\Delta^{-2}\sum_{i=1}^{k}\left(\widehat{\Sigma}_i+\rho I\right)^{-2}\widehat{\Sigma}_i-\Delta^{-2}\sum_{i=1}^{k}\left(\widehat{\Sigma}_i+\rho I\right)^{-3}\widehat{\Sigma}_i\right]\\
&=\operatorname{tr}\left[\Delta^{-2}\left(\sum_{i=1}^{k}\left(\widehat{\Sigma}_i+\rho I\right)^{-2}\widehat{\Sigma}_i\cdot\Delta^{-1}\sum_{i=1}^{k}\left(\widehat{\Sigma}_i+\rho I\right)^{-2}\widehat{\Sigma}_i-\sum_{i=1}^{k}\left(\widehat{\Sigma}_i+\rho I\right)^{-3}\widehat{\Sigma}_i\right)\right]\leq 0
\end{aligned}$$

Let $A=\sum_{i=1}^{k}\left(\widehat{\Sigma}_i+\rho I\right)^{-3}\widehat{\Sigma}_i$, $B=\sum_{i=1}^{k}\left(\widehat{\Sigma}_i+\rho I\right)^{-2}\widehat{\Sigma}_i$, then we only need to show

$$\operatorname{tr}\left[\Delta^{-2}\left(A-B^\top\Delta^{-1}B\right)\right]\geq 0.$$

Since $\Delta^{-2}$ is a positive semi-definite matrix, and for two positive semi-definite matrices $X=P^\top P, Y=Q^\top Q$ we have $\operatorname{tr}(XY)=\operatorname{tr}(P^\top PQ^\top Q)=\operatorname{tr}(PQ^\top QP^\top)\geq 0$, the remaining work is to show $A-B^\top\Delta^{-1}B\succeq 0$. Now, by using Lemma 2.8.6, it enough to show

$$M=\begin{bmatrix}A & B\\ B^\top & \Delta\end{bmatrix}\succeq 0.$$

We can write $M$ as sum of matrices $M_i$, where

$$M_i=\begin{bmatrix}\left(\widehat{\Sigma}_i+\rho I\right)^{-3}\widehat{\Sigma}_i & \left(\widehat{\Sigma}_i+\rho I\right)^{-2}\widehat{\Sigma}_i\\ \left(\widehat{\Sigma}_i+\rho I\right)^{-2}\widehat{\Sigma}_i & \left(\widehat{\Sigma}_i+\rho I\right)^{-1}\widehat{\Sigma}_i\end{bmatrix}.$$

By using Lemma 2.8.6 again, $M_i\succeq 0$ since

$$\left(\widehat{\Sigma}_i+\rho I\right)^{-3}\widehat{\Sigma}_i-\left(\widehat{\Sigma}_i+\rho I\right)^{-2}\widehat{\Sigma}_i\cdot\left(\widehat{\Sigma}_i+\rho I\right)\widehat{\Sigma}_i^{-1}\cdot\left(\widehat{\Sigma}_i+\rho I\right)^{-2}\widehat{\Sigma}_i=0.$$

Finally we have $M=\sum_{i=1}^{k}M_i\succeq 0$, i.e. $\psi'(\rho)\leq 0$. The special case when $\rho=0$ is of special

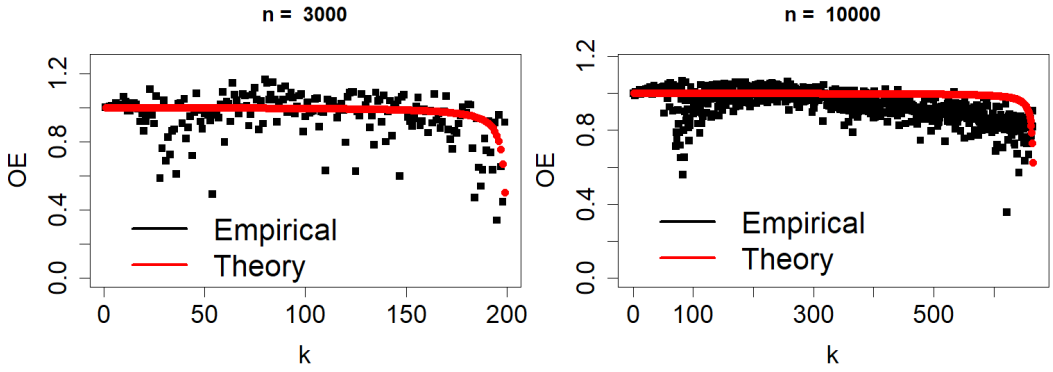



Figure 7: NYC flights data.

interest. In this case, $\Delta = kI$ and we can simplify the derivative

$$\psi'(0) = \frac{2}{nk^2} \cdot \operatorname{tr}\left[\left(\sum_{i=1}^{k} \widehat{\Sigma}_i^{-1}\right)^2 - k\sum_{i=1}^{k} \widehat{\Sigma}_i^{-2}\right].$$

By using the Cauchy-Schwarz inequality for positive semidefinite matrices

$$\operatorname{tr}(AB) \leq \sqrt{\operatorname{tr}(A^2)\operatorname{tr}(B^2)} \leq \frac{\operatorname{tr}(A^2) + \operatorname{tr}(B^2)}{2},$$

we can easily verify that $\psi'(0) \leq 0$, and equality holds if and only if $X_1^\top X_1 = X_2^\top X_2 = \cdots = X_k^\top X_k$.

*2.8.18. Empirical data analysis*

In this section we present an empirical data example to assess the accuracy of our theoretical results. Specifically, figure 7 shows a comparison of our theoretical formulas for OE and actual out-of-sample prediction error (test error) on the NYC flights dataset (Wickham, 2018). We observe a quite good match.

Specifically, we performed the following steps in our data analysis. We downloaded the flights data as included in the nycflights13 R package (Wickham, 2018). We joined the separate datasets (weather, planes, and airlines). We omitted data points with missing entries. We removed one out of each pair of variables with absolute correlation higher than

0.8. This left a total of $N = 60,448$ samples and $p = 17$ variables. For $n = 3000, 10000$, we randomly sampled a training set of size $n$, and a non-overlapping test set of size also equal to $n$. The test set size does not have equal the training set size, and we only followed this protocol for simplicity.

We then fit linear regression estimators to this data in a global and distributed way. For the distributed version, we split the train data as equally as possible into $k$ subsets, for each $k \leq n/p$. We then fit a linear regression to each subset, and took a weighted average with the optimal weights. We computed the test error of both the global and the distributed estimators over the test sample, and defined their ratio to be the empirical OE. We compared this to our theoretical formula for the OE, see figure 7.

We observe a quite good match between the theoretical and empirical results. However, the empirical estimate of OE can be larger than unity. This is because of sampling noise. Our results show that $OE \leq 1$, but only for the theoretical quantity where we have taken expectations. To get estimators with reduced variance, one could average over multiple Monte Carlo trials. However, those are beyond the our scope.

# CHAPTER 3 : WONDER: Weighted One-shot Distributed Ridge Regression in High Dimensions

This chapter is based on Dobriban and Sheng (2019), which is a joint work with my advisor Professor Edgar Dobriban. I contributed to a large portion of ideas, derivations, and simulations.

## 3.1. Introduction

Computers have changed all aspects of our world. Importantly, computing has made data analysis more convenient than ever before. However, computers also pose limitations and challenges for data science. For instance, hardware architecture is based on a model of a universal computer—a Turing machine—but in fact has physical limitations of storage, memory, processing speed, and communication bandwidth over a network. As large data sets become more and more common in all areas of human activity, we need to think carefully about working with these limitations.

How can we design methods for data analysis (statistics and machine learning) that scale to large data sets? A general approach is *distributed and parallel computing*. Roughly speaking, the data is divided up among computing units, which perform most of the computation locally, and synchronize by passing relatively short messages. While the idea is simple, a good implementation can be hard and nontrivial. Moreover, different problems have different inherent needs in terms of local computation and global communication resources. For instance, in statistical problems with high levels of noise, simple one-shot schemes like averaging estimators computed on local data sets can sometimes work well.

In this paper, we study a fundamental problem in this area. We are interested in linear regression, which is arguably one of the most important problems in statistics and machine learning. A popular method for this model is *ridge regression* (aka Tikhonov regularization), which regularizes the estimates using a quadratic penalty to improve estimation and prediction accuracy. We aim to understand how to do ridge regression in a distributed computing

environment. We are also interested in the important *high-dimensional* setting, where the number of features can be very large. In fact our approach allows the dimension and sample size to have any ratio. We also work in a random-effects model where each predictor has a small effect on the outcome, which is the model for which ridge regression is best suited.

We consider the simplest and most fundamental method, which performs ridge regression locally on each data set housed on the individual machines or other computing units, sends the estimates to a global datacenter (or parameter server), and then constructs a final one-shot estimator by taking a linear combination of the local estimates. As mentioned, such methods are sometimes near-optimal, and it is therefore well-justified to study them. We will later give several additional justifications for our work.

However, in contrast to existing work, we introduce a completely new mathematical approach to the problem, which has never been used for studying distributed ridge regression before. Specifically, we leverage and further develop sophisticated recent techniques from random matrix theory and free probability theory in our analysis. This enables us to make important contributions, that were simply unattainable using more "traditional" mathematical approaches.

To give a sense of our results, we provide a brief discussion here. We have a data set consisting of $n$ datapoints, for instance 1000 heart disease patients. Each datapoint has an outcome $y_j$, such as blood pressure, and features $x_j$, such as age, height, electronic health records, lab results, and genetic variables. Our goal is to predict the outcome of interest (i.e., blood pressure) for new patients based on their features, and to estimate the relationship of the outcome to the features.

The samples are distributed across several sites, for instance patients from different countries are housed in different data centers. We will refer to the sites as "machines", though they may actually be other computing entities, such as entire computer networks or data centers. In many important settings, it can be impossible to share the data across the different sites, for instance due to logistical or privacy reasons.

Therefore, we assume that each site has a subset of the samples. Our approach is to train ridge regression on this local data. As usual, we can arrange the local data set (say on the $i$-th machine) into a feature matrix $X_i$, where each row contains a sample (i.e., datapoint), and an outcome vector $Y_i$ where each entry is an outcome. We compute the local ridge regression estimates

$$\hat{\beta}_i = (X_i^\top X_i + \lambda_i I_p)^{-1} X_i^\top Y_i,$$

where $\lambda_i$ are some regularization parameters. We then aggregate them by a weighted combination, constructing the final *one-shot distributed ridge estimator* (where $k$ is the number of sites)

$$\hat{\beta}_{dist} = \sum_{i=1}^{k} w_i \hat{\beta}_i.$$

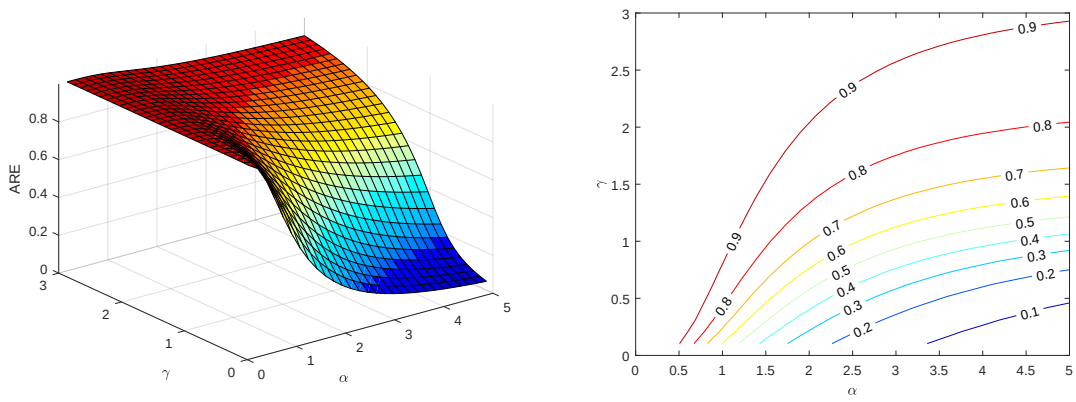
The important questions here are:

1. How does this work?

2. How to tune the parameters? (such as the regularization parameters and weights)

Question 1 is of interest because we wish to know when one-shot methods are a good approach, and when they are not. For this we need to understand the performance as a function of the key problem parameters, such as the signal strength, sample size, and dimension. For question 2, the challenge is posed by the constraints of the distributed computing environment, where standard methods for parameter tuning such as cross-validation may be expensive.

In this work we are able to make several crucial contributions to these questions. We work in an asymptotic setting where $n, p$ grow to infinity at the same rate, which effectively gives good results for any $n, p$. We study a linear-random effects model, where each regressor has a small random effect on the outcome. This is a good model for the applications where ridge regression is used, because ridge does not assume sparsity, and has optimality properties in certain dense random effects models. Importantly, this analysis does *not* assume any




(a) Surface

(b) Contour

Figure 8: Efficiency loss due to one-shot distributed learning. This plot shows the relative mean squared error of centralized ridge regression compared to optimally weighted one-shot distributed ridge regression. This quantity is at most unity, and the larger, the "better" distributed ridge works. Specifically, the model is asymptotic, and we show the dependence of the Asymptotic Relative Efficiency (ARE) on the aspect ratio $\gamma = \lim p/n$ (where $n$ is sample size and $p$ is dimension) and on the signal strengh $\alpha = \sqrt{\mathbb{E}\|\beta\|^2}$, in the *infinite-worker limit* when we distribute our data over many machines. We show (a) surface and (b) contour plots of the ARE. See the text for details.

sparsity in a high-dimensional setting. Sparsity has been one of the biggest driving forces in statistics and machine learning in the last 20 years. Our work is in a different line of work, and shows that meaningful results are available without sparsity.

We find the limiting mean squared error of the one-shot distributed ridge estimator. This enables us to characterize the optimal weights and tuning parameters, as well as the *relative efficiency* compared to centralized ridge regression, meaning the ratio of the risk of usual ridge to the distributed estimator. This can precisely pinpoint the computation-accuracy tradeoff achieved via one-shot distributed estimation. See Figure 8 for an illustration.

As a consequence of our detailed and precise risk analysis, we make several qualitative discoveries that we find quite striking:

1. **Efficiency depends strongly on signal strength.** The statistical efficiency of the one-shot distributed ridge estimator depends strongly on signal strength. The efficiency is

generally high (meaning distributed ridge regression works well) when the signal strength is low.

2. **Infinite-worker limit.** The one-shot distributed estimator does not lose all efficiency compared to the ridge estimator even in the limit of *infinitely many machines*. Somewhat surprisingly, this suggests that simple one-shot weighted combination methods for distributed ridge regression can work well even for very large numbers of machines. The statement that this can be achieved by communication-efficient methods is nontrivial. This finding is clearly important from a practical perspective.

3. **Decoupling.** When the features are uncorrelated, the problem of choosing the optimal regularization parameters *decouples* over the different machines. We can choose them in a locally optimal way, and they are also globally optimal. We emphasize that this is a very delicate result, and is not true in general for correlated features. Moreover, this discovery is also important in practice, because it gives conditions under which we can choose the regularization parameters separately for each machine, thus saving valuable computational resources.

4. **Optimal weights do not sum to unity.** Our work uncovers unexpected properties of the optimal weights. Naively, one may think that the weights need to sum to unity, meaning that we need a weighted average. However, it turns out the optimal weights sum to more than unity, because of the negative bias of the ridge estimator. This means that *any type of averaging method is suboptimal*. We characterize the optimal weights and under certain conditions find their explicit analytic form.

Based on these results, we propose a new Weighted ONe-shot DistributEd Ridge regression algorithm (WONDER). We also confirm these results in detailed simulation studies and on an empirical data example, using the Million Song Dataset. Here WONDER can be used over 100-way splits of the data with 5% loss of prediction accuracy.

We also emphasize that some aspects of our work can help practitioners directly (e.g., our

new algorithm), while others are developed for deepening our understanding of the nature of the problem. We discuss the practical implications of our work in Section 3.4.5.

The paper is structured as follows. We discuss some related work in Section 3.1.1. We start with finite sample results in Section 3.2. We provide asymptotic results for features with an arbitrary covariance structure in Section 3.3. We consider the special case of an identity covariance in Section 3.4. In Section 3.5 we provide an explicit algorithm for optimally weighted one-shot distributed ridge. We also study in detail the properties of the estimation error, relative efficiency (including minimax properties in Section 3.4.6), tuning parameters (and decoupling), as well as optimal weights, including answers to the questions above. We provide numerical simulations throughout the paper, and additional ones in Section 3.6, along with an example using an empirical data set.

*3.1.1. Related work*

Here we discuss some related work. Historically, distributed and parallel computation has first been studied in computer science and optimization (see e.g., Bertsekas and Tsitsiklis, 1989; Lynch, 1996; Blelloch and Maggs, 2010; Boyd et al., 2011; Rauber and Rünger, 2013; Koutris et al., 2018). However, the problems studied there are quite different from the ones that we are interested in. Those works often focus on problems where correct answers are required within numerical precision, e.g., 16 bits of accuracy. However, when we have noisy data sets, such as in statistics and machine learning, numerical precision is neither needed nor usually possible. We may only hope for 3-4 bits of accuracy, and thus the problems are different.

The area of distributed statistics and machine learning has attracted increasing attention only relatively recently, see for instance Mcdonald et al. (2009); Zhang et al. (2012, 2013b); Li et al. (2013); Zhang et al. (2013a); Duchi et al. (2014); Chen and Xie (2014); Mackey et al. (2011); Zhang et al. (2015); Braverman et al. (2016); Jordan et al. (2016); Rosenblatt and Nadler (2016); Smith et al. (2016); Banerjee et al. (2019a); Zhao et al. (2016); Xu et al. (2018); Fan et al. (2017); Lin et al. (2017); Lee et al. (2017); Volgushev et al. (2019a); Shang

and Cheng (2017); Battey et al. (2018); Zhu and Lafferty (2018); Chen et al. (2019, 2018c); Wang et al. (2019); Shi et al. (2018); Duan et al. (2018); Liu et al. (2018b); Cai and Wei (2020), and the references therein. See Huo and Cao (2018) for a review. We can only discuss the most closely related papers due to space limitations.

Zhang et al. (2013b) study the MSE of averaged estimation in empirical risk minimization. Later Zhang et al. (2015) study divide and conquer *kernel ridge regression*, showing that the partition-based estimator achieves the statistical minimax rate over all estimators, when the number of machines is not too large. These results are very general, however they are not as explicit or precise as our results. In addition they consider fixed dimensions, whereas we study increasing dimensions under random effects models. Lin et al. (2017) improve the above results, removing certain eigenvalue assumptions on the kernel, and sharpening the rate.

Guo et al. (2017) study regularization kernel networks, and propose a debiasing scheme that can improve the behavior of distributed estimators. This work is also in the same framework as those above (general kernel, fixed dimension). Xu et al. (2018) propose a distributed General Cross-Validation method to choose the regularization parameter.

Rosenblatt and Nadler (2016) consider averaging in distributed learning in fixed and high-dimensional M-estimation, without studying regularization. Lee et al. (2017) study sparse linear regression, showing that averaging debiased lasso estimators can achieve the optimal estimation rate if the number of machines is not too large. A related work is Battey et al. (2018), which also includes hypothesis testing under more general sparse models. These last two works are on a different problem (sparse regression), whereas we study ridge regression in random-effects models.

## 3.2. Finite Sample Results

We start our study of distributed ridge regression by a finite sample analysis of estimation error in linear models. Consider the standard linear model

$$Y = X\beta + \varepsilon. \tag{3.1}$$

Here $Y \in \mathbb{R}^n$ is the $n$-dimensional continuous outcome vector of $n$ independent samples (e.g., the blood pressure level of $n$ patients, or the amount of time spent on an activity by $n$ internet users), $X$ is the $n \times p$ design matrix containing the values of $p$ features for each sample (e.g., demographical and genetic variables of each patient). Moreover, $\beta = (\beta_1, \ldots, \beta_p)^\top \in \mathbb{R}^p$ is the $p$-dimensional vector of unknown regression coefficients.

Our goals are to predict the outcome variable for future samples, and also to estimate the regression coefficients. The outcome vector is affected by the random noise $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^\top \in \mathbb{R}^n$. We assume that the coordinates of $\varepsilon$ are independent random variables with mean zero and variance $\sigma^2$.

The *ridge regression* (or Tikhonov regularization) estimator is one of the most popular methods for estimation and prediction in linear models. Recall that the ridge estimator of $\beta$ is

$$\hat{\beta}(\lambda) = (X^\top X + n\lambda I_p)^{-1} X^\top Y,$$

where $\lambda$ is a tuning parameter. This estimator has many justifications. It shrinks the coefficients of the usual ordinary least squares estimator, which can lead to improved estimation and prediction. When the entries of $\beta$ and $\varepsilon$ are i.i.d. Gaussian, and for suitable $\lambda$, it is the posterior mean of $\beta$ given the outcomes, and hence is a Bayes optimal estimator for any quadratic loss function, including estimation and prediction error.

Suppose now that we are in a distributed computation setting. The samples are distributed across $k$ different sites or machines. For instance, the data of users from a particular country may be stored in a separate datacenter. This may happen due to memory or

storage limitations of individual data storage facilities, or may be required by data usage agreements. As mentioned, for simplicity we call the sites "machines".

We can write the partitioned data as

$$X = \begin{bmatrix} X_1 \\ \vdots \\ X_k \end{bmatrix}, \; Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_k \end{bmatrix}.$$

Thus the $i$-th machine contains $n_i$ samples whose features are stored in the $n_i \times p$ matrix $X_i$ and also the corresponding $n_i \times 1$ outcome vector $Y_i$.

Since the ridge regression estimator is a widely used gold standard method, we would like to understand how we can approximate it in a distributed setting. Specifically, we will focus on one-shot *weighting* methods, where we perform ridge regression locally on each subset of the data, and then aggregate the regression coefficients by a weighted sum. There are several reasons to consider weighting methods:

1. This is a practical method with *minimal communication cost*. When communication is expensive, it is imperative to develop methods that minimize communication cost. In this case, one-shot weighting methods are attractive, and so it is important to understand how they work. In a well-known course on scalable machine learning, Alex Smola calls such methods "idiot-proof" (Smola, 2012), meaning that they are straightforward to implement (unlike some of the more sophisticated methods).

2. Averaging (which is a special case of one-shot weighting) has already been studied in several works on distributed ridge regression (e.g., Zhang et al. (2015); Lin et al. (2017)), and much more broadly in distributed learning, see the related work section for details. Such methods are *known to be rate-optimal* under certain conditions.

3. However, in our setting, we are able to discover several *new phenomena* about one-shot weighting. For instance, we can quantify in a much more nuanced way the accuracy loss

compared to centralized ridge regression.

4. Weighting may serve as a useful *initialization to iterative methods*. In practical distributed learning problems, iterative optimization algorithms such as distributed gradient descent or ADMM (Boyd et al., 2011) may be used. However, there are examples where the first step of the iterative method has *worse* performance than a simple averaging (Pourshafeie et al., 2018). Therefore, we can imagine hybrid or warm start methods that use weighting as an initialization. This also suggests that studying one-shot weighting is important.

Therefore, we define *local* ridge estimators for each data set $X_i, Y_i$, with regularization parameter $\lambda_i$ as

$$\hat{\beta}_i(\lambda_i) = (X_i^\top X_i + n_i \lambda_i I_p)^{-1} X_i^\top Y_i.$$

We consider combining the local ridge estimators at a central server via a one-step weighted summation. We will find the optimally weighted one-shot distributed estimator

$$\hat{\beta}_{dist}(w) = \sum_{i=1}^{k} w_i \hat{\beta}_i.$$

Note that, unlike ordinary least squares (OLS), the local ridge estimators are always well-defined, i.e. $n_i$ can be smaller than $p$. Also, for the distributed OLS estimator averaging local OLS solutions, it is natural to require $\sum_i w_i = 1$, because this ensures unbiasedness (Dobriban and Sheng, 2018). However, the ridge estimators are biased, so it is not clear if we should put any constraints on the weights. In fact we will find that the optimal weights typically do not sum to unity. These features distinguish our work from prior art, and lead to some surprising consequences.

Throughout the paper, we will frequently use the notations $\widehat{\Sigma} = n^{-1} X^\top X$ and $\widehat{\Sigma}_i = n_i^{-1} X_i^\top X_i$. A stepping stone to our analysis is the following key result.

**Theorem 3.2.1** (Finite sample risk and optimal weights)**.** *Consider the distributed ridge regression problem described above. Suppose we have a data set with n datapoints (samples),*

*each with an outcome and p features. The data set is distributed across k sites. Each site has a subset $X_i, Y_i$ of the data, with the $n_i \times p$ matrix $X_i$ of features of $n_i$ samples, and the corresponding outcomes $Y_i$. We compute the local ridge regression estimator $\hat{\beta}_i(\lambda_i) = (X_i^\top X_i + n_i\lambda_i I_p)^{-1} X_i^\top Y_i$ with fixed regularization parameters $\lambda_i > 0$ on each data set. We send the local estimates to a central location, and combine them via a weighted sum, i.e., $\hat{\beta}_{dist}(w) = \sum_{i=1}^k w_i \hat{\beta}_i$.*

*Under the linear regression model* (3.1)*, the optimal weights that minimize the mean squared error of the distributed estimator are*

$$w^* = (A + R)^{-1} v,$$

*where the quantities $v, A, R$ are defined below.*

1. *$v$ is a $k$-dimensional vector with $i$-th coordinate $\beta^\top Q_i \beta$, where $Q_i = (\widehat{\Sigma}_i + \lambda_i I_p)^{-1} \widehat{\Sigma}_i$ are $p \times p$ matrices.*

2. *$A$ is a $k \times k$ matrix with $(i, j)$-th entry $\beta^\top Q_i Q_j \beta$.*

3. *$R$ is a $k \times k$ diagonal matrix with $i$-th diagonal entry $n_i^{-1} \sigma^2 \operatorname{tr}[(\widehat{\Sigma}_i + \lambda_i I_p)^{-2} \widehat{\Sigma}_i]$.*

*The mean squared error of the* optimally weighted *distributed ridge regression estimator $\hat{\beta}_{dist}$ with $k$ sites equals*

$$\mathrm{MSE}^*_{dist}(k) = \mathbb{E}\|\hat{\beta}_{dist}(w^*) - \beta\|^2 = \|\beta\|^2 - v^\top (A + R)^{-1} v,$$

See Appendix 3.7.6 for the proof. The argument proceeds via a direct calculation, recognizing that *finding the optimal weights for combining the local estimators $\hat{\beta}_i$ can be viewed as a $k$-parameter regression problem* of $\beta$ on $\hat{\beta}_i$, for $i = 1, \ldots, k$.

This result quantifies the mean squared error of the optimally weighted distributed ridge estimator *for fixed regularization parameters $\lambda_i$*. Later we will study how to choose the

regularization parameters optimally. The result also gives an exact formula for the optimal weights. However, the optimal weights depend on the unknown regression coefficients $\beta$, and are thus not directly usable in practice. Instead, our approach is to make stronger assumptions on $\beta$ under which we can develop estimators for the weights.

**Computational efficiency.** We take a short detour here to discuss computational efficiency. Here by computational efficiency we mean the total time consumption. Computing one ridge regression estimator $(X^\top X + \lambda I_p)^{-1} X^\top Y$ for a fixed regularization parameter $\lambda$ and $n \times p$ design matrix $X$ can be done in time $O(np \min(n, p))$ by first computing the SVD of $X$. This automatically gives the ridge estimator for all values of $\lambda$.

How much time can we save by distributing the data? Suppose first that $n \geq p$, in which case the total time consumption is $O(np^2)$. Computing ridge locally on the $i$-th machine takes $O(n_i p \min(n_i, p))$ time. Suppose next that we distribute equally to $k$ of machines, and we also have $n_i = n/k \geq p$. Then the time consumption is reduced to $O((n/k)p^2) = O(np^2/k)$. In this case we can say that the total time consumption decreases *proportionally to the number of machines*. This shows the benefit of parallel data processing.

On the other extreme, if $n \leq p$, then $n_i = n/k \leq p$, the total time consumption is reduced from $O(n^2 p)$ to $O((n/k)^2 p) = O(n^2 p/k^2)$. This shows that the total time consumption decreases *quadratically* in the number of machines (albeit of course the constant is much worse). If we are in an intermediate case where $n \geq p$ and $n_i = n/k \leq p$, then the time decreases at a rate between linear and quadratic.

*3.2.1. Addressing reader concerns*

At this stage, our readers may have several concerns about our approach. We address some concerns in turn below.

1. Does it make sense to average ridge estimators, which can be biased?

   A possible concern is that we are working with biased estimators. Would it make sense to debias them first, before weighting? A similar approach has been used for sparse

regression, with the debiased Lasso estimators (Lee et al., 2017; Battey et al., 2018). However, our results *allow the regularization parameters to be arbitrarily close to zero*, which leads to least squares estimators, with an inverse or pseudoinverse $(X_i^\top X_i)^\dagger$. These are the "natural" debiasing estimators for ridge regression. For OLS, these are exactly unbiased, while for pseudoinverse, they are approximately so. Hence our approach allows nearly unbiased estimators, and we automatically discover when this is the optimal method.

2. Is it possible to improve the weighted sum of local ridge estimators $\hat{\beta}_i$ in trivial ways?

   One-shot weighting is merely a heuristic. If it were possible to improve it in a simple way, then it would make sense to study those methods instead of weighting. However, we are not aware of such methods. For instance, one possibility is to try and add the constant vector into the regression on the global parameter server, because this may help reduce the bias. In simulation studies, we have observed that this approach does not usually lead to a perceptible decrease in MSE. Specifically we have found that under the simulation setting common throughout the paper, the MSEs with and without a constant term are close (see Appendix 3.7.1 for details).

## 3.3. Asymptotics under Linear Random-effects Models

The finite sample results obtained so far can be hard to interpret, and do not allow us to directly understand the performance of the optimal one-shot distributed estimator. Therefore, we will consider an asymptotic setting that leads to more insightful results.

Recall that our basic linear model is $Y = X\beta + \varepsilon$, where the error $\varepsilon$ is random. Next, we also assume that a *random-effects model* holds. We assume $\beta$ is random—independently of $\varepsilon$—with coordinates that are themselves independent random variables with mean zero and variance $p^{-1}\sigma^2\alpha^2$. Thus, each feature contributes a small random amount to the outcome. Ridge regression is designed to work well in such a setting, and has several optimality properties in variants of this model. The parameters are now $\theta = (\sigma^2, \alpha^2)$: the *noise level* $\sigma^2$ and the *signal-to-noise ratio* $\alpha^2$ respectively. This parametrization is standard and

widely used (e.g. Searle et al. (2009); Dicker and Erdogdu (2017); Dobriban and Wager (2018)).

To get more insight into the performance of ridge regression in a distributed environment, we will take an asymptotic approach. Notice from Theorem 3.2.1 that the mean squared error depends on the data only through simple functionals of the sample covariance matrices $\widehat{\Sigma}$ and $\widehat{\Sigma}_i$, such as

$$\beta^\top(\widehat{\Sigma}_i + \lambda_i I_p)^{-1}\widehat{\Sigma}_i\beta, \quad \beta^\top(\widehat{\Sigma}_i + \lambda_i I_p)^{-1}\widehat{\Sigma}_i(\widehat{\Sigma}_j + \lambda_j I_p)^{-1}\widehat{\Sigma}_j\beta, \quad \mathrm{tr}[(\widehat{\Sigma}_i + \lambda_i I_p)^{-2}\widehat{\Sigma}_i].$$

When the coordinates of $\beta$ are i.i.d., the means of the quadratic functionals become proportional to the *traces* of functions of the sample covariance matrices. This motivates us to adopt models from *asymptotic random matrix theory*, where the asymptotics of such quantities are a central topic.

We begin by introducing some key concepts from random matrix theory (RMT) which will be used in our analysis. We will focus on "Marchenko-Pastur" (MP) type sample covariance matrices, which are fundamental and popular in statistics (see e.g., Bai and Silverstein (2010); Anderson (2003); Paul and Aue (2014); Yao et al. (2015)). A key concept is the spectral distribution, which for a $p \times p$ symmetric matrix $A$ is the distribution $F_A$ that places equal mass on all eigenvalues $\lambda_i(A)$ of $\Sigma$. This has cumulative distribution function (CDF) $F_A(x) = p^{-1}\sum_{i=1}^p \mathbf{1}(\lambda_i(A) \leq x)$. A central result in the area is the Marchenko-Pastur theorem, which states that eigenvalue distributions of sample covariance matrices converge (Marchenko and Pastur, 1967; Bai and Silverstein, 2010). We state the required assumptions below:

**Assumption 2.** *Consider the following conditions:*

1. *The $n \times p$ design matrix $X$ is generated as $X = Z\Sigma^{1/2}$ for an $n \times p$ matrix $Z$ with i.i.d. entries (viewed as coming from an infinite array), satisfying $\mathbb{E}[Z_{ij}] = 0$ and $\mathbb{E}[Z_{ij}^2] = 1$, and a deterministic $p \times p$ positive semidefinite population covariance matrix $\Sigma$.*

2. *The sample size $n$ grows to infinity proportionally with the dimension $p$, i.e. $n, p \to \infty$ and $p/n \to \gamma \in (0, \infty)$.*

3. *The sequence of spectral distributions $F_\Sigma := F_{\Sigma,n,p}$ of $\Sigma := \Sigma_{n,p}$ converges weakly to a limiting distribution $H$ supported on $[0, \infty)$, called the population spectral distribution.*

Then, the Marchenko-Pastur theorem states that with probability 1, the spectral distribution $F_{\widehat{\Sigma}}$ of the sample covariance matrix $\widehat{\Sigma}$ also converges weakly (in distribution) to a limiting distribution $F_\gamma := F_\gamma(H)$ supported on $[0, \infty)$ (Marchenko and Pastur, 1967; Bai and Silverstein, 2010). The limiting distribution is determined uniquely by a fixed-point equation for its *Stieltjes transform*, which is defined for any distribution $G$ supported on $[0, \infty)$ as

$$m_G(z) := \int_0^\infty \frac{1}{t - z} dG(t), \quad z \in \mathbb{C} \setminus \mathbb{R}^+.$$

With this notation, the Stieltjes transform of the spectral measure of $\widehat{\Sigma}$ satisfies

$$m_{\widehat{\Sigma}}(z) = p^{-1} \operatorname{tr}[(\widehat{\Sigma} - zI_p)^{-1}] \to_{a.s.} m_{F_\gamma}(z), \quad z \in \mathbb{C} \setminus \mathbb{R}^+,$$

where $m_{F_\gamma}(z)$ is the Stieltjes transform of $F$. In addition, we denote by $m'(z)$ the derivative of the Stieltjes transform. Then, it is also known that

$$p^{-1} \operatorname{tr}[(\widehat{\Sigma} - zI_p)^{-2}] \to_{a.s.} m'_{F_\gamma}(z).$$

The results stated above can be expressed in a different, and perhaps slightly more modern language, using *deterministic equivalents* (Serdobolskii, 2007; Hachem et al., 2007; Couillet et al., 2011; Dobriban and Sheng, 2018). For instance, the Marchenko-Pastur law is a consequence of the following result. For any $z$ where it is well-defined, consider the resolvent $(\widehat{\Sigma} - zI_p)^{-1}$. This random matrix is *equivalent* to a deterministic matrix $(x_p\Sigma - zI_p)^{-1}$ for

a certain scalar $x_p = x(\Sigma, n, p, z)$, and we write

$$(\widehat{\Sigma} - zI_p)^{-1} \asymp (x_p\Sigma - zI_p)^{-1}.$$

Here two sequences of $n \times n$ matrices $A_n, B_n$ (not necessarily symmetric) of growing dimensions are *equivalent*, and we write

$$A_n \asymp B_n$$

if

$$\lim_{n\to\infty} \operatorname{tr}\left[C_n(A_n - B_n)\right] = 0$$

almost surely, for any sequence $C_n$ of $n \times n$ deterministic matrices (not necessarily symmetric) with bounded trace norm, i.e., such that $\limsup \|C_n\|_{tr} < \infty$ (Dobriban and Sheng, 2018). Informally, any linear combination of the entries of $A_n$ can be approximated by the entries of $B_n$. This also can be viewed as a kind of *weak convergence* in the matrix space equipped with an inner product (trace). From this, it also follows that the traces of the two matrices are equivalent, from which we can recover the MP law.

In Dobriban and Sheng (2018), we collected some useful properties of the calculus of deterministic equivalents. In this work, we use those properties extensively. We also develop and use a new *differentiation rule for the calculus of deterministic equivalents* (see Appendix 3.7.2).

We are now ready to study the asymptotics of the risk. We express the limits of interest in *two equivalent forms*, one in terms of *population quantities* (such as the limiting spectral distribution $H$ of $\Sigma$), and one in terms of *sample quantities* (such as the limiting spectral distribution $F_\gamma$ of $\widehat{\Sigma}$). Moreover, we will denote by $T$ a random variable distributed according to $H$, so that $\mathbb{E}_H[g(T)]$ denotes the mean of $g(T)$ when $T$ is a random variable distributed according to the limit spectral distribution $H$.

The key to obtaining the results based on population quantities is that the quadratic

forms involving $\beta$ have asymptotic equivalents that only depend on $\alpha^2, \sigma^2$, based on the concentration of quadratic forms. Specifically, we have

$$\beta^\top A \beta \approx \frac{1}{p}\sigma^2\alpha^2 \cdot \mathrm{tr}(A)$$

for suitable matrices $A$ (see the proof of Theorem 3.3.1 for details). The key to the results based on sample quantities is the MP law and the calculus of deterministic equivalents.

**Theorem 3.3.1** (Asymptotics for distributed ridge, arbitrary subsample size)**.** *In the linear random-effects model under Assumption 2, suppose in addition that the eigenvalues of $\Sigma$ are uniformly bounded away from zero and infinity, and that the entries of $Z$ have a finite $(8+\varepsilon)$-th moment for some $\varepsilon > 0$. Suppose moreover that the local sample sizes $n_i$ grow proportionally to $p$, so that $p/n_i \to \gamma_i > 0$.*

*Then the optimal weights for distributed ridge regression, and its mean square error, converge to definite limits. Recall from Theorem 3.2.1 that we have the formulas $w^* = (A+R)^{-1}v$ and $MSE^*_{dist} = \|\beta\|^2 - v^\top(A+R)^{-1}v$ for the optimal finite sample weights and risk, and thus it is enough to find the limit of $v, A$ and $R$. These have the following limits:*

1. *With probability one, we have the convergence $v \to V \in \mathbb{R}^k$. The $i$-th coordinate of the limit $V$ has the following two equivalent forms, in terms of population and sample quantities, respectively:*

$$V_i = \sigma^2\alpha^2\mathbb{E}_H\frac{x_i T}{x_i T + \lambda_i} = \sigma^2\alpha^2\left[1 - \lambda_i m_{F_{\gamma_i}}(-\lambda_i)\right].$$

*Recall that $H$ is the limiting population spectral distribution of $\Sigma$, and $T$ is a random variable distributed according to $H$. Among the empirical quantities, $F_{\gamma_i}$ is the limiting empirical spectral distribution of $\widehat{\Sigma}_i$ and $x_i := x_i(H, \lambda_i, \gamma_i) > 0$ is the unique solution of the fixed point equation*

$$1 - x_i = \gamma_i\left[1 - \lambda_i\int_0^\infty \frac{dH(t)}{x_i t + \lambda_i}\right] = \gamma_i\left[1 - \mathbb{E}_H\frac{\lambda_i}{x_i T + \lambda_i}\right].$$

*It is part of the theorem's claim that there is such an $x_i$.*

2. *With probability one, $A \to \mathcal{A} \in \mathbb{R}^{k\times k}$. For $i \neq j$, the $(i,j)$-th entry of $\mathcal{A}$ is, in terms of the population spectral distribution $H$,*

$$\mathcal{A}_{ij} = \sigma^2\alpha^2\mathbb{E}_H \frac{x_i x_j T^2}{(x_i T + \lambda_i)(x_j T + \lambda_j)}.$$

*The $i$-th diagonal entry of $\mathcal{A}$ is, in terms of population and sample quantities, respectively,*

$$\begin{aligned}\mathcal{A}_{ii} &= \sigma^2\alpha^2\left[1 - \mathbb{E}_H \frac{2\lambda_i x_i T + \lambda_i^2}{(x_i T + \lambda_i)^2} + \frac{\lambda_i^2\gamma_i x_i \left(\mathbb{E}_H \frac{T}{(x_i T+\lambda_i)^2}\right)^2}{1 + \gamma_i\lambda_i \mathbb{E}_H \frac{T}{(x_i T+\lambda_i)^2}}\right] \\ &= \sigma^2\alpha^2\left[1 - 2\lambda_i m_{F_{\gamma_i}}(-\lambda_i) + \lambda_i^2 m'_{F_{\gamma_i}}(-\lambda_i)\right].\end{aligned}$$

3. *With probability one, the diagonal matrix $R$ converges, $R \to \mathcal{R} \in \mathbb{R}^{k\times k}$, where of course $\mathcal{R}$ is also diagonal. The $i$-th diagonal entry of $\mathcal{R}$ is, in terms of population and sample quantities, respectively,*

$$\mathcal{R}_{ii} = \sigma^2\left[\frac{x_i \mathbb{E}_H \frac{T}{(x_i T+\lambda_i)^2}}{1 + \lambda_i\gamma_i \mathbb{E}_H \frac{T}{(x_i T+\lambda_i)^2}}\right] = \sigma^2\left[\gamma_i m_{F_{\gamma_i}}(-\lambda_i) - \gamma_i\lambda_i m'_{F_{\gamma_i}}(-\lambda_i)\right].$$

*The limiting weights and mean square error are then*

$$\mathcal{W}_k^* = (\mathcal{A} + \mathcal{R})^{-1} V$$

*and*

$$\mathcal{M}_k = \sigma^2\alpha^2 - V^\top(\mathcal{A} + \mathcal{R})^{-1} V.$$

See Appendix 3.7.7 for the proof. The statement may look complicated, but the formulas simplify considerably in the uncorrelated case $\Sigma = I_p$, on which we will focus later. Moreover, these limiting formulas are also fundamental for developing consistent estimators

for the optimal weights. To develop an algorithm for the practically common general covariance case, the following theorem is crucial.

**Theorem 3.3.2** (Asymptotics for distributed ridge, equal subsample size)**.** *Consider the assumptions and the notations of Theorem 3.3.1. We further assume the samples are equally distributed across the local machines, i.e.* $n_1 = n_2 = \cdots = n_k = n/k$ *and* $\gamma_1 = \gamma_2 = \cdots = \gamma_k = k\gamma$*. We use the same tuning parameter* $\lambda$ *for each local estimator. Then the limiting optimal weights* $\mathcal{W}_k^*$ *and the limiting MSE* $\mathcal{M}_k$ *have the following forms:*

$$\mathcal{W}_k^* = (1, 1, \ldots, 1)^\top \cdot \frac{\sigma^2\alpha^2(1-\lambda m)}{\mathcal{F} + k\mathcal{G}} \quad \text{and} \quad \mathcal{M}_k = \sigma^2\alpha^2 - \frac{\sigma^4\alpha^4(1-\lambda m)^2 k}{\mathcal{F} + k\mathcal{G}}.$$

*Here* $\mathcal{F}$ *and* $\mathcal{G}$ *are defined as follows:*

$$\mathcal{F} = \sigma^2\alpha^2 \frac{k\gamma\lambda^2(m-\lambda m')^2}{1 - k\gamma + k\gamma\lambda m'} + \sigma^2 k\gamma(m - \lambda m')$$

*and*

$$\mathcal{G} = \sigma^2\alpha^2 \left(1 - 2\lambda m + \lambda^2 m' - \frac{k\gamma\lambda^2(m-\lambda m')^2}{1 - k\gamma + k\gamma\lambda m'}\right)$$

*where* $m := m_{F_{k\gamma}}(-\lambda)$ *and* $m' := -\frac{dm}{d\lambda}$.

See Appendix 3.7.8 for the proof and an explanation of why we need to assume the samples are uniformly distributed. Based on this theorem, we are able to develop an algorithm which works for arbitrary covariance structures. See Section 3.5 for the details.

Now we discuss the problem of estimating the optimal weights, which is crucial for developing practical methods. The results in Theorem 3.3.2 show that to estimate the weights consistently, if the tuning parameter $\lambda$ is known, we *only need to estimate* $\alpha^2, \sigma^2$ consistently. The reason is that we can use $\mathrm{tr}[(\widehat{\Sigma}_i + \lambda I)^{-1}]/p$ to approximate $m$, and use $\mathrm{tr}[(\widehat{\Sigma}_i + \lambda I)^{-2}]/p$ to approximate $m'$.

Estimating these two parameters is a well-known problem, and several approaches have been proposed, for instance restricted maximum likelihood (REML) estimators (Jiang, 1996;

Searle et al., 2009; Dicker, 2014b; Dicker and Erdogdu, 2016; Jiang et al., 2016), etc. We can use—for instance—results from Dicker and Erdogdu (2017), who showed that the Gaussian MLE is consistent and asymptotically efficient for $\theta = (\sigma^2, \alpha^2)$ even in the non-Gaussian setting of this paper (see Appendix 3.7.3 for a summary).

## 3.4. Special Case: Identity Covariance

When the population covariance matrix is the identity, that is $\Sigma = I$, the results simplify considerably. In this case the features are nearly uncorrelated. It is known that the limiting Stieltjes transform $m_{F_\gamma} := m_\gamma$ of $\widehat{\Sigma}$ has the explicit form (Marchenko and Pastur, 1967):

$$m_\gamma(z) = \frac{(z+\gamma-1) + \sqrt{(z+\gamma-1)^2 - 4z\gamma}}{-2z\gamma}. \tag{3.2}$$

As usual in the area, we use the principal branch of the square root of complex numbers.

### *3.4.1. Properties of the estimation error and asymptotic relative efficiency*

We can use the closed form expression for the Stieltjes transform to get explicit formulas for the optimal weights. From Theorem 3.3.1, we conclude the following simplified result:

**Theorem 3.4.1** (Asymptotics for isotropic population covariance)**.** *In addition to the assumptions of Theorem 3.3.1, suppose that the population covariance matrix $\Sigma = I$. Then the limits of $v, A$ and $R$ have simple explicit forms:*

1. *The $i$-th coordinate of $V$ is:*

$$V_i = \sigma^2\alpha^2 \left[1 - \lambda_i m_{\gamma_i}(-\lambda_i)\right],$$

*where $m_{\gamma_i}(-\lambda_i)$ is the Stieltjes transform given above in equation* (3.2)*.*

2. *The entries of $\mathcal{A}$ are*

$$\mathcal{A}_{ij} = \begin{cases} \sigma^2\alpha^2 \left[1 - \lambda_i m_{\gamma_i}(-\lambda_i)\right] \cdot \left[1 - \lambda_j m_{\gamma_j}(-\lambda_j)\right], & \text{for } i \neq j \\ \sigma^2\alpha^2 \left[1 - 2\lambda_i m_{\gamma_i}(-\lambda_i) + \lambda_i^2 m'_{\gamma_i}(-\lambda_i)\right], & \text{for } i = j. \end{cases}$$

*3. The i-th diagonal entry of* $\mathcal{R}$ *is*

$$\mathcal{R}_{ii} = \sigma^2 \gamma_i \left[ m_{\gamma_i}(-\lambda_i) - \lambda_i m'_{\gamma_i}(-\lambda_i) \right].$$

*The limiting optimal weights for combining the local ridge estimators are* $\mathcal{W}_k^* = (\mathcal{A}+\mathcal{R})^{-1}V$*, and MSE of the optimally weighted distributed estimator is*

$$\mathcal{M}_k = \frac{\sigma^2 \alpha^2}{1 + \sum_{i=1}^k \frac{V_i^2}{\sigma^2\alpha^2(\mathcal{R}_{ii}+\mathcal{A}_{ii})-V_i^2}}.$$

See Appendix 3.7.9 for the proof. This theorem shows the surprising fact that the limiting risk *decouples* over the different machines. By this we mean that the limiting risk can be written in a simple form, involving a sum of terms depending on each machine, without any interaction. This seems like a major surprise.

To explain in more detail the decoupling phenomenon, let us study how the local risks are related to the distributed risks. Define $V = V(\gamma, \lambda)$ to be the limiting scalar $V \in \mathbb{R}$ defined above, in the special case $k = 1$. Explicitly, this is the limit of the quantity $\beta^\top Q \beta$, where $Q = (\widehat{\Sigma} + \lambda I_p)^{-1}\widehat{\Sigma}$, as given in Theorem 3.2.1 applied for $k = 1$. Let $D$ be the scalar expression $D(\gamma, \lambda) = \sigma^2\alpha^2(\mathcal{R} + \mathcal{A}) - V$ when $k = 1$. With these notations, the risk $\mathcal{M}_1$ of ridge regression when computed on the entire data set equals

$$\mathcal{M}_1(\gamma, \lambda) = \frac{\sigma^2 \alpha^2}{1 + \frac{V(\gamma,\lambda)}{D(\gamma,\lambda)}}.$$

Moreover, the risk of optimally weighted one-shot distributed ridge over $k$ subsets, with arbitrary regularization parameters $\lambda_i$, equals

$$\mathcal{M}_k(\gamma_1, \ldots, \gamma_k, \lambda_1, \ldots, \lambda_k) = \frac{\sigma^2 \alpha^2}{1 + \sum_{i=1}^k \frac{V_i^2(\gamma_i,\lambda_i)}{D_i(\gamma_i,\lambda_i)}}.$$

Then one can check that we have the following equations connecting the risk computed on

the entire data set and the distributed risk:

$$\frac{\sigma^2\alpha^2}{\mathcal{M}_k(\gamma_1,\ldots,\gamma_k,\lambda_1,\ldots,\lambda_k)} - 1 = \sum_{i=1}^{k} \frac{\sigma^2\alpha^2}{\mathcal{M}_1(\gamma_i,\lambda_i)} - k,$$
$$\mathcal{M}_k(\gamma_1,\ldots,\gamma_k,\lambda_1,\ldots,\lambda_k) = \frac{1}{\sum_{i=1}^{k} \frac{1}{\mathcal{M}_1(\gamma_i,\lambda_i)} + \frac{1-k}{\sigma^2\alpha^2}}.$$

These equations are precisely what we mean by *decoupling*. The distributed risk can be written as a function of the type $1/(\sum_i 1/x_i + b)$ of the distributed risks. Therefore, there are no "interactions" between the different risk functions. Similar expressions have been obtained for linear regression (Dobriban and Sheng, 2018).

Next, we discuss in more depth why the limiting risk decouples. Mathematically, the key reason is that when $\Sigma = I$, the limit of $A_{ij}$ for $i \neq j$ decouples into a product of two terms. Therefore, the distributed risk function involves a quadratic form with zero *off-diagonal* terms. This is not the case for general population covariance $\Sigma$. We provide an explanation via free probability theory in Appendix 3.7.4.

An important consequence of the decoupling is that *we can optimize the individual risks over the tuning parameters $\lambda_i$ separately.*

**Proposition 3.4.2** (Optimal regularization (tuning) parameters)**.** *Under the assumptions of Theorem 3.4.1, the optimal regularization (tuning) parameters $\lambda_i$ that minimize the local MSEs also minimize the distributed risk $\mathcal{M}_k$. They have the form*

$$\lambda_i = \frac{\gamma_i}{\alpha^2}, \quad i = 1, 2, \ldots, k.$$

*Moreover, the risk $\mathcal{M}_k$ of distributed ridge regression with optimally tuned regularization parameters is*

$$\mathcal{M}_k = \frac{\sigma^2\alpha^2}{1 + \sum_{i=1}^{k} \left[\frac{\alpha^2}{\gamma_i m_{\gamma_i}(-\gamma_i/\alpha^2)} - 1\right]},$$

See Appendix 3.7.10 for the proof.

The main goal of our paper is to study the behavior of the one-shot distributed ridge estimator and compare it with the centralized estimator. It is helpful to first understand the properties of the *optimal risk* function $\phi(\gamma) := \gamma m_\gamma(-\gamma/\alpha^2)$. The optimal risk function equals the optimally tuned global risk $\mathcal{M}_1$ up to a factor $\sigma^2$. It has the explicit form

$$\phi(\gamma) = \gamma m_\gamma(-\gamma/\alpha^2) = \frac{-\gamma/\alpha^2 + \gamma - 1 + \sqrt{(-\gamma/\alpha^2 + \gamma - 1)^2 + 4\gamma^2/\alpha^2}}{2\gamma/\alpha^2}.$$

**Proposition 3.4.3** (Properties of the optimal risk function)**.** *The optimal risk function $\phi(\gamma)$ has the following properties:*

1. **Monotonicity:** *$\phi(\gamma)$ is an increasing function of $\gamma \in [0, \infty)$ with $\lim_{\gamma \to 0+} \phi(\gamma) = 0$ and $\lim_{\gamma \to +\infty} \phi(\gamma) = \alpha^2$.*

2. **Concavity:** *When $\alpha \leq 1, \phi(\gamma)$ is a concave function of $\gamma \in [0, \infty)$. When $\alpha > 1$, $\phi(\gamma)$ is convex for small $\gamma$ (close to 0), and concave for large $\gamma$.*

See Appendix 3.7.11 for the proof. See also Figure 9 for plots of $\phi$ for different $\alpha$, which show its monotonicity and convexity properties. The aspect ratio $\gamma$ characterizes the dimensionality of the problem. It makes sense that $\phi(\gamma)$ is increasing, since the regression problem should become more difficult as the dimension increases. For the second property, the concavity of the function means that it grows very fast to approach its limit. When the signal-to-noise ratio $\alpha^2$ is small, the risk is concave, so it grows fast with the dimension. But when the signal-to-noise ratio becomes large, the risk will grow much slower at the beginning. Here the phase transition happens at $\alpha^2 = 1$. This gives insight into the effect of the signal-to-noise ratio on the regression problem.

To compare the distributed and centralized estimators, we will study their (asymptotic) relative efficiency (ARE), which is the (limit of the) ratio of their mean squared errors. Here we assume each estimator is optimally tuned. This quantity, which is at most unity, captures the loss of efficiency due to the distributed setting. An ARE close to 1 is "good",

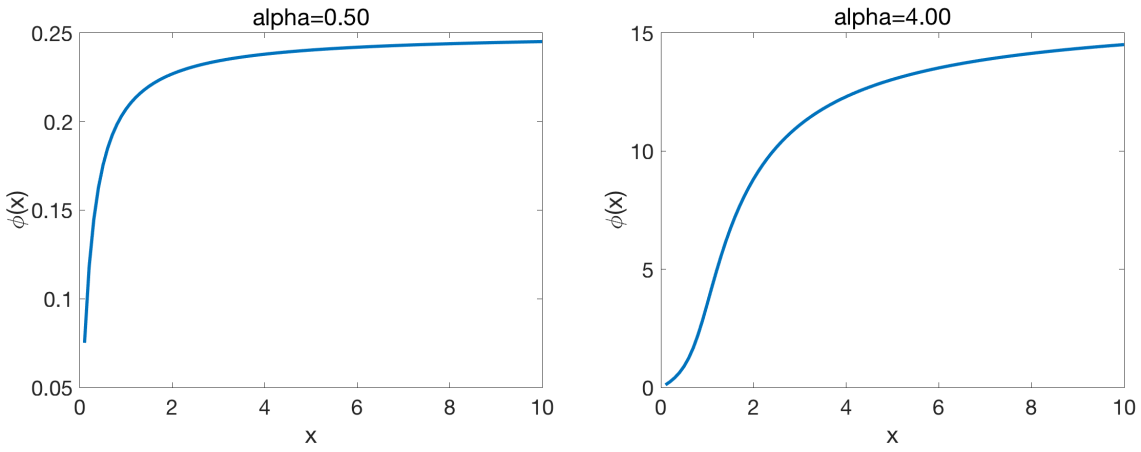



Figure 9: Plots of the optimal risk function $\phi$ as a function of the aspect ratio $\gamma$ (denoted by $x$ in the plots), for different signal strength parameters $\alpha$.

while an ARE close to 0 is "bad". From the results above, it follows that the ARE has the form

$$ARE = \frac{\mathcal{M}_1}{\mathcal{M}_k} = \frac{\gamma m_\gamma(-\gamma/\alpha^2)}{\alpha^2}\left[1 + \sum_{i=1}^{k}\left(\frac{\alpha^2}{\gamma_i m_{\gamma_i}(-\gamma_i/\alpha^2)} - 1\right)\right] \leq 1.$$

We have the following properties of the ARE.

**Theorem 3.4.4** (Properties of the asymptotic relative efficiency (ARE))**.** *The asymptotic relative efficiency (ARE) has the following properties:*

1. **Worst case is equally distributed data***: For fixed $k, \alpha^2$ and $\gamma$, the ARE attains its minimum when the samples are* equally distributed *across $k$ machines, i.e. $\gamma_1 = \gamma_2 = \cdots = \gamma_k = k\gamma$. We denote the minimal value by $\psi(k, \gamma, \alpha^2)$. That is*

$$\min_{\gamma_1,\ldots,\gamma_k} ARE = \psi(k, \gamma, \alpha^2) := \frac{\gamma m_\gamma(-\gamma/\alpha^2)}{\alpha^2}\left(1 - k + \frac{\alpha^2}{\gamma m_{k\gamma}(-k\gamma/\alpha^2)}\right).$$

2. **Adding more machines leads to efficiency loss***: For fixed $\alpha^2$ and $\gamma$, $\psi(k, \gamma, \alpha^2)$ is a decreasing function on $k \in [1, \infty)$ with $\lim_{k\to 1+} \psi(k, \gamma, \alpha^2) = 1$ and* infinite-worker limit

$$\lim_{k\to\infty} \psi(k,\gamma,\alpha^2) = h(\alpha^2,\gamma) < 1.$$

*Here we can view $\psi$ as a continuous function of $k$ for convenience, although originally it is only well-defined for $k \in \mathbb{N}$. We emphasize that the infinite-worker limit tells us how much efficiency we have for a very large number of machines. It is a nontrivial result that this quantity is strictly positive.*

3. **Form of the infinite-worker limit**: *As a function of $\alpha^2$ and $\gamma$, $h(\alpha^2,\gamma)$ has the explicit form*

$$h(\alpha^2,\gamma) = \frac{-\gamma/\alpha^2+\gamma-1+\sqrt{(-\gamma/\alpha^2+\gamma-1)^2+4\gamma^2/\alpha^2}}{2\gamma}\left(1+\frac{\alpha^2}{\gamma(1+\alpha^2)}\right).$$

4. **Edge cases of the infinite-worker limit**: *For fixed $\alpha^2$, $h(\alpha^2,\gamma)$ is an increasing function of $\gamma \in [0,\infty)$ with limit*

$$\lim_{\gamma\to 0} h(\alpha^2,\gamma) = \frac{1}{1+\alpha^2}, \quad \lim_{\gamma\to\infty} h(\alpha^2,\gamma) = 1.$$

*On the other hand, for fixed $\gamma$, $h(\alpha^2,\gamma)$ is a decreasing function of $\alpha^2 \in [0,\infty)$ with limit*

$$\lim_{\alpha^2\to 0} h(\alpha^2,\gamma) = 1, \quad \lim_{\alpha^2\to\infty} h(\alpha^2,\gamma) = \begin{cases} 1-\frac{1}{\gamma^2}, & \gamma > 1, \\ 0, & 0 < \gamma \le 1. \end{cases}$$

See Appendix 3.7.12 for the proof. See Figure 10 for some plots of the evenly distributed ARE $\psi$ for various $\alpha$ and $\gamma$ and Figure 8 for the surface and contour plots of $h(\alpha^2,\gamma)$. The efficiency loss tends to be larger (ARE is smaller) when the signal-to-noise ratio $\alpha^2$ is larger. The plots confirm the theoretical result that the efficiency always decreases with the number of machines. Relatively speaking, the distributed problem becomes easier and easier as the dimension increases, compared to the aggregated problem (i.e., the ARE increases in $\gamma$ for

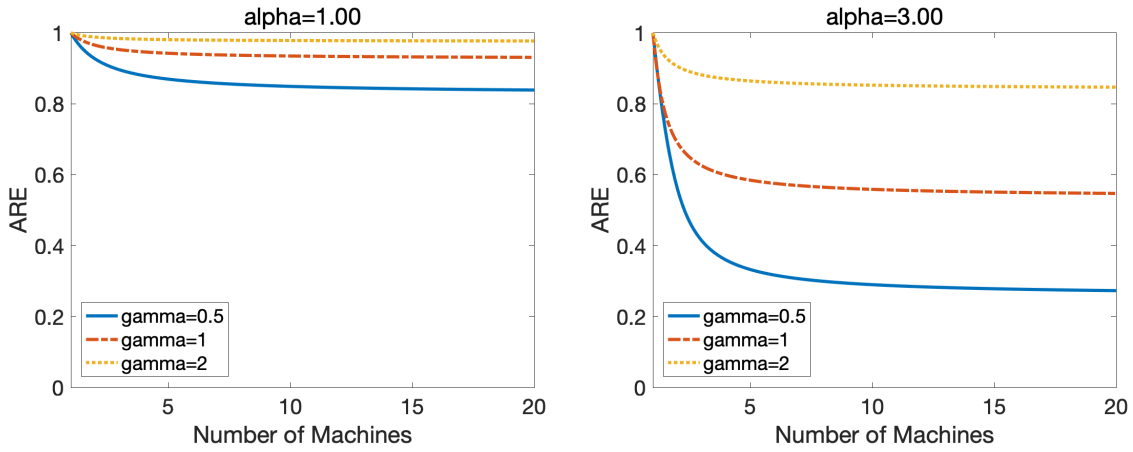



Figure 10: Plots of the asymptotic relative efficiency $\psi$ when the data set are evenly distributed, for different $\alpha$ and $\gamma$. See Theorem 3.4.4 for the properties of the ARE.

fixed parameters). This can be viewed as a blessing of dimensionality.

We also observe a nontrivial *infinite-worker limit*. Even in the limit of many machines, distributed ridge *does not lose all efficiency*. This is in contrast to doing linear regression on each machine, where all efficiency is lost when the local sample sizes are less than the dimension (Dobriban and Sheng, 2018). This is one of the few results in the distributed learning literature where one-step weighting gives nontrivial results for *arbitrary large* $k$, i.e., we can take $k \to \infty$ and we still obtain nontrivial results. We find this quite remarkable.

Overall, the ARE is generally large, *except* when $\gamma$ is small and $\alpha$ is large. This is a setting with strong signal and relatively low dimension, which is also the "easiest" setting from a statistical point of view. In this case, perhaps we should use other techniques for distributed estimation, such as iterative methods.

*3.4.2. Properties of the optimal weights*

Next, we study properties of the optimal weights. This is important, because choosing them is a crucial practical question. The literature on distributed regression typically considers simple averages of local estimators, for which $\hat{\beta}_{dist} = k^{-1} \sum_{i=1}^{k} \hat{\beta}_i$ (see, e.g. Zhang et al. (2015); Lee et al. (2017); Battey et al. (2018)). In contrast, we will find that the optimal

weights *do not sum up to unity.*

Formally, we have the following properties of the optimal weights.

**Theorem 3.4.5** (Properties of the optimal weights)**.** *The asymptotically optimal weights* $\mathcal{W}_k^* = (\mathcal{A} + \mathcal{R})^{-1} V$ *have the following properties:*

1. **Form of the optimal weights***: The* $i$*-th coordinate of* $\mathcal{W}_k$ *is:*

$$\mathcal{W}_{k,i} = \left( \frac{\alpha^2}{\gamma_i m_{\gamma_i}(-\gamma_i/\alpha^2)} \right) \cdot \left( \frac{1}{1 + \sum_{i=1}^{k} \left[ \frac{\alpha^2}{\gamma_i m_{\gamma_i}(-\gamma_i/\alpha^2)} - 1 \right]} \right),$$

   *and the sum of the limiting weights is always greater than or equal to one:*

$$\sum_{i=1}^{k} \mathcal{W}_{k,i} \geq 1.$$

   *When* $k \geq 2$*, the sum is strictly greater than one.*

2. **Evenly distributed optimal weights***: When the samples are evenly distributed, so that all limiting aspect ratios* $\gamma_i$ *are equal,* $\gamma_i = k\gamma$*, then all* $\mathcal{W}_{k,i}$ *equal the* optimal weight function $\mathcal{W}(k, \gamma, \alpha^2)$*, which has the form*

$$\mathcal{W}(k, \gamma, \alpha^2) = \frac{\alpha^2}{\alpha^2 k + (1-k) k\gamma \cdot m_{k\gamma}(-k\gamma/\alpha^2)}.$$

   *This can also be written in terms of the* optimal risk function $\phi(\gamma, \alpha^2)$ *defined above as*

$$\mathcal{W}(k, \gamma, \alpha^2) = \frac{\alpha^2}{\alpha^2 k - (k-1)\phi(k\gamma, \alpha^2)}.$$

3. **Limiting cases***: For fixed* $k$ *and* $\alpha^2$*, the optimal weight function* $\mathcal{W}(k, \gamma, \alpha^2)$ *is an increasing function of* $\gamma \in [0, \infty)$ *with* $\lim_{\gamma \to 0+} \mathcal{W}(\gamma) = 1/k$ *and* $\lim_{\gamma \to \infty} \mathcal{W}(\gamma) = 1$*.*

See Appendix 3.7.13 for the proof. See Figures 11 and 12 for some plots of the optimal weight function with $k = 2$. We can see that the optimal weights are usually large, and always

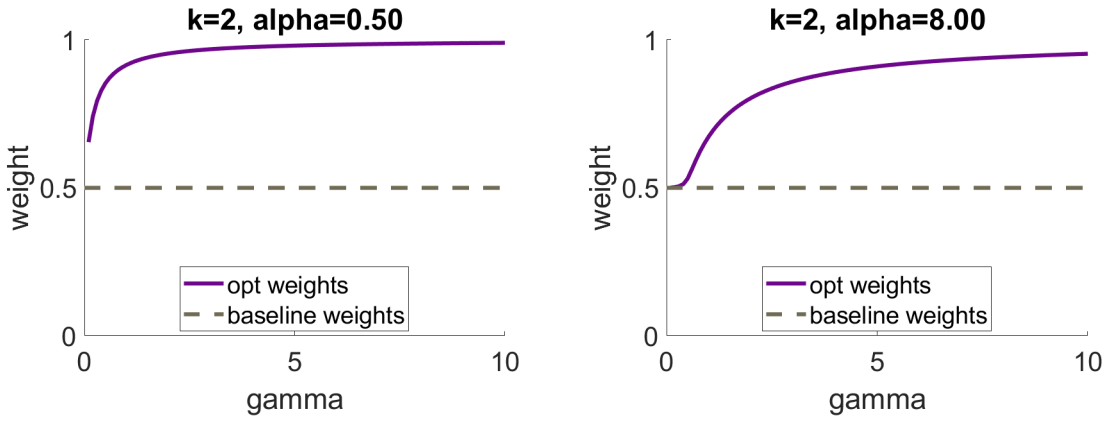
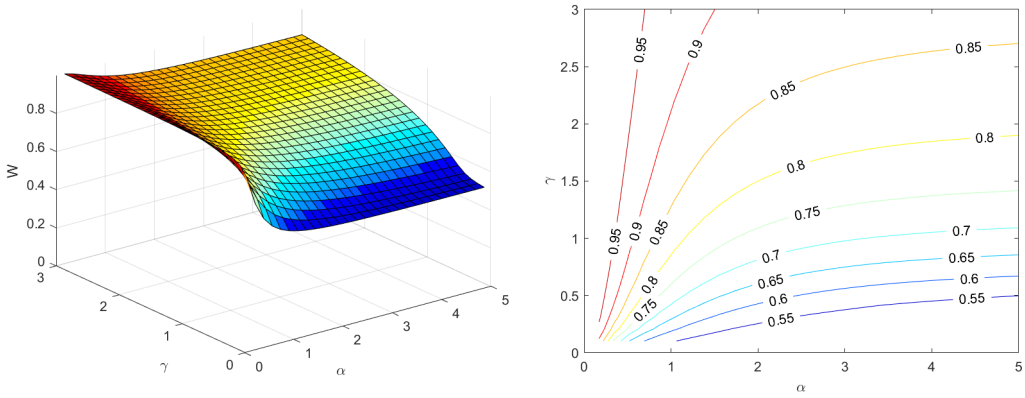



Figure 11: Plots of optimal weights for different $\alpha$.




(a) Surface (b) Contour

Figure 12: Surface and contour plots of the optimal weights.

greater than $1/k$. When the signal-to-noise ratio $\alpha^2$ is small, the weight function is concave and increases fast to approach one. In the low dimensional setting where $\gamma \to 0$, the weights tend to the uniform average $1/k$. Hence in this setting we recover the classical uniform averaging methods, which makes sense, because ridge regression with optimal regularization parameter tends to linear regression in this regime.

How much does optimal weighting help? It is both interesting and important to know this, especially compared to naive uniform weighting, because it allows us to compare our proposed weighting method to the "baseline". See Figure 13. We have plotted the risk of distributed ridge regression for both the optimally weighted version and the simple average,

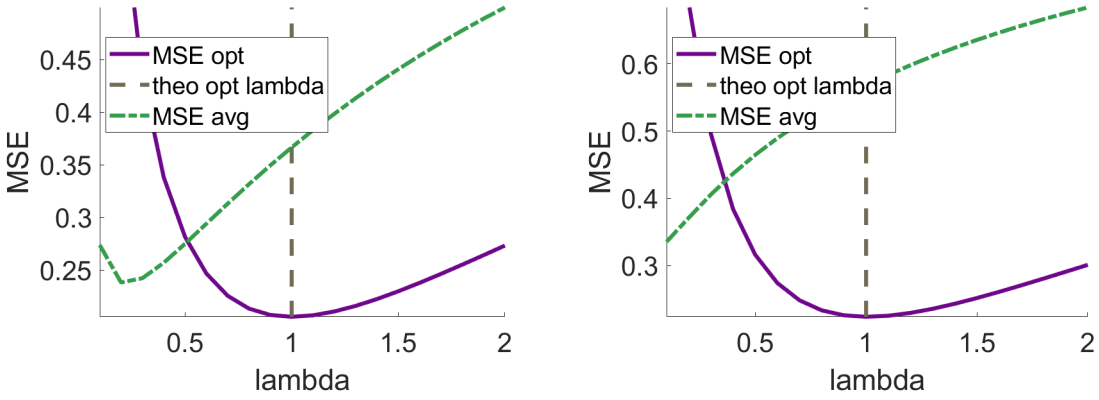



Figure 13: Distributed risk as a function of the regularization parameter. We plot both the risk with optimal weights (MSE opt) and the risk obtained from sub-optimal averaging (MSE avg). We set $\alpha = 1$, $\gamma = 0.17$ and $k = 5, 10$.

as a function of the regularization parameter. We observe that *optimal weighting can lead to a 30-40% decrease in the risk*. Therefore, our proposed weighting scheme can lead to a substantial benefit.

Why are the weights large, and why do they sum to a quantity greater than one? The short intuitive answer is that ridge regression is *negatively* (or *downward*) biased, and so we must *counter the effect of bias by upweighting*. This also can be viewed as a way of *debiasing*. In different contexts, it is already well known that debiasing can play a kew role in distributed learning (Lee et al. (2017); Battey et al. (2018)). We provide a slightly more detailed intuitive explanation in Appendix 3.7.5.

*3.4.3. Out-of-sample prediction*

So far, we have discussed the estimation problem. In real applications, out-of-sample prediction is also of interest. We consider a test datapoint $(x_t, y_t)$, generated from the same model $y_t = x_t^\top \beta + \varepsilon_t$, where $x_t, \varepsilon_t$ are independent of $X, \varepsilon$. We want to use $x_t^\top \hat{\beta}$ to predict $y_t$, and the out-of-sample prediction error is defined as $\mathbb{E}(y_t - x_t^\top \hat{\beta})^2$. Then we have the following proposition.

**Proposition 3.4.6** (Out-of-sample prediction error and relative efficiency)**.** *Under the conditions of Theorem 3.4.1, the limiting out-of-sample prediction error of the optimal*

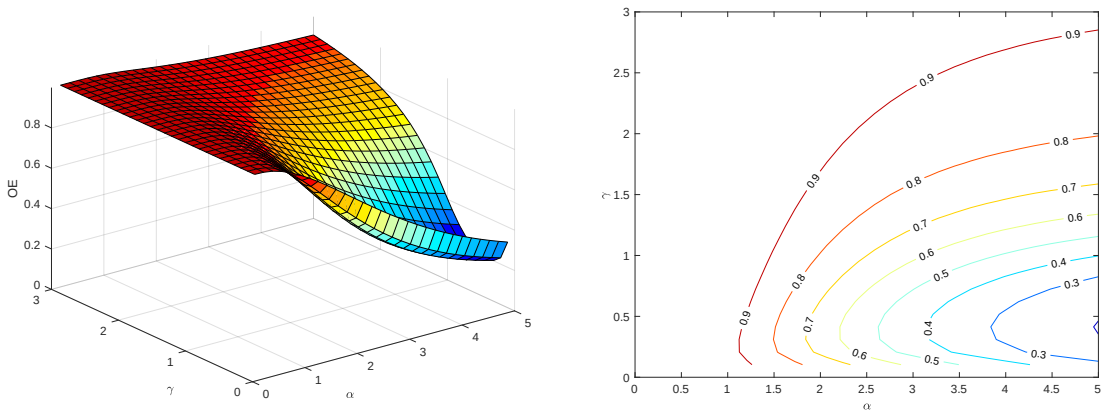



Figure 14: Limit of OE: (a) surface and (b) contour plots of $\mathcal{H}(\alpha^2, \gamma)$.

*distributed estimator* $\hat{\beta}_{dist}$ *is*

$$\mathcal{O}_k = \sigma^2 + \mathcal{M}_k.$$

*Thus, the asymptotic out-of-sample relative efficiency, meaning the ratio of prediction errors, is*

$$OE = \frac{\mathcal{O}_1}{\mathcal{O}_k} = \frac{\mathcal{M}_1 + \sigma^2}{\mathcal{M}_k + \sigma^2},$$

*and the efficiency for prediction is higher than for estimation* $OE \geq ARE$. *Furthermore, when the samples are equally distributed, the relative efficiency has the form*

$$\Psi(k, \gamma, \alpha^2) = \frac{1 + \gamma m_\gamma(-\gamma/\alpha^2)}{1 + \frac{\alpha^2 \gamma m_{k\gamma}(-k\gamma/\alpha^2)}{\alpha^2 + (1-k)\gamma m_{k\gamma}(-k\gamma/\alpha^2)}},$$

*and the corresponding infinite-worker limit (taking* $k \to \infty$*) is*

$$\mathcal{H}(\alpha^2, \gamma) = \frac{1 + \gamma m_\gamma(-\gamma/\alpha^2)}{1 + \frac{\gamma\alpha^2(1+\alpha^2)}{\alpha^2 + \gamma(1+\alpha^2)}}.$$

See Appendix 3.7.14 for the proof and Figure 14 for some plots. This proposition implies that, for the identity covariance case, the efficiency loss of the distributed estimator in terms of the test error is always less than the loss in terms of the estimation error. When the signal-to-noise ratio $\alpha^2$ is small, the relative efficiency is always very large and close to 1.

This observation can be an encouragement to use our distributed methods for out-of-sample prediction.

*3.4.4. Choosing the regularization parameter*

Previous work found that, under certain conditions, the regularization parameters on the individual machines should be chosen as if they had the all samples (Zhang et al., 2015). Our findings are consistent with these results. However, the reasons behind our findings are very different from prior work. The intuition for the previous results is that the *variance* of distributed estimators averages out, while the *bias* does not do so. Therefore, the regularization parameters should be chosen such that the local bias is lower than for locally optimal tuning. This means that we should use smaller regularization parameters locally.

In our case, we find that for isotropic covariance, the optimal risk *decouples across machines.* Hence, the regularization parameters on the machines can be chosen optimally for each machine. Moreover, in our asymptotics the locally optimal choice is a *constant multiple* of the globally optimal choice, namely the multiple in front of the identity matrix in the local ridge estimator $(X_i^\top X_i + n_i\lambda_i I_p)^{-1} X_i^\top Y_i$ should be $\lambda_i = p/(n_i\alpha^2)$ whereas the globally optimal $\lambda$ is $\lambda = p/(n\alpha^2)$.

Roughly speaking, this derivation reaches the same conclusion as prior work about the choice of regularization parameters, namely that the regularization parameters on the machines should be chosen as if they had the all samples. However, we emphasize that our results are very different, because the optimal weighting procedure has weights summing to *greater than unity.* Moreover, we also consider the proportional-limit case, and the conclusion for regularization parameters only applies to the isotropic case.

*3.4.5. Implications and practical relevance*

We discuss some of the implications of our results. Our finite-sample results show that the optimal way to weight the estimators depends on functionals of the unknown parameter $\beta$, while the asymptotic results in general depend on the eigenvalues of $\widehat{\Sigma}$ (or $\Sigma$). These are unavailable in practice, and hence these results can typically not be used on real data

sets. However, since our results are precise and accurate (they capture the *truth* about the problem), we interpret this as saying that *the problem is hard in general*. Meaning that optimal weighting for ridge regression is a challenging statistical problem. In practice that means that we may be content with uniform weighting. It remains to be investigated in future work how much we should up-adjust those equal weights.

The optimal weights become usable in the case of spherical data, when $\Sigma = I$ (or, more accurately, the limiting spectral distribution of $\Sigma$ is the point mass at unity). In practice, we can get closer to this assumption by using some form of *whitening* on the data, for instance by scaling all variables to the same scale, by estimating $\Sigma$ over restricted classes, such as assuming block-covariance structures. Alternatively, we can use correlation screening, where we remove features with high correlation. At this stage, all these approaches are heuristic, but we include them to explain how our results can be relevant in practice. It is a topic of future research to make these ideas more concrete. In the algorithm we proposed in Section 3.5, we use grid search to find a good tuning parameter under general covariance structures.

On the theoretical side, our results can also be interpreted as a form of *reduction* between statistical problems. *If* we can estimate the quadratic functionals of the unknown regression parameter involved in our weights, *then* we can do optimally weighted ridge regression. In this sense, we reduce distributed ridge regression to the estimation of those quadratic functionals. We think that in the challenging and novel setting of distributed learning, such reductions can be both interesting and potentially useful.

An important question is "Should we use distributed linear or ridge regression?". If we have $n_i \geq p$ and linear regression is defined on each local machine, then we can use either distributed linear (Dobriban and Sheng, 2018) or ridge regression. Linear regression has the advantage that the optimal weights are easy to find. Therefore, if we cannot reasonably reduce to the case $\Sigma = I$, it seems we should use linear regression.

*3.4.6. Minimax optimality of the optimal distributed estimator*

To deepen our understanding of the distributed problem, we next show that the optimal distributed ridge estimator is asymptotically rate-minimax. Suppose without loss of generality that the noise level $\sigma^2 = 1$, and let $\mathbb{S}^{p-1}(\alpha) = \{\beta \in \mathbb{R}^p; ||\beta|| = \alpha\}$ denote the sphere of radius $\alpha \geq 0$ in $\mathbb{R}^p$ centered at the origin. Then the minimax risk for estimating $\beta$ over the sphere $\mathbb{S}^{p-1}(\alpha)$ is

$$r(\alpha) = \inf_{\hat{\beta}} \sup_{\beta \in \mathbb{S}^{p-1}(\alpha)} R(\hat{\beta}, \beta) = \inf_{\hat{\beta}} \sup_{\beta \in \mathbb{S}^{p-1}(\alpha)} \mathbb{E}_\beta ||\hat{\beta} - \beta||^2,$$

where the expectation is over both $X$ and $\varepsilon$. This problem has been well studied by Dicker (2014a), who reduced it to the following Bayes problem. Let $\pi$ be the uniform measure on $\mathbb{S}^{p-1}(\alpha)$. Then the Bayes risk with respect to $\pi$ is

$$r_B(\alpha) = \inf_{\hat{\beta}} \int_{\mathbb{S}^{p-1}(\alpha)} R(\hat{\beta}, \beta) d\pi(\beta) = \inf_{\hat{\beta}} \mathbb{E}_\pi ||\hat{\beta} - \beta||^2.$$

The Bayes estimator is the posterior mean $\hat{\beta}_{\mathbb{S}^{p-1}(\alpha)} = \mathbb{E}_\pi(\beta | y, X)$. So the corresponding Bayes risk is $r_B(\alpha) = \mathbb{E}_\pi ||\hat{\beta}_{\mathbb{S}^{p-1}(\alpha)} - \beta||^2$. Then, the Bayes estimator also minimizes the original minimax risk and $r(\alpha) = r_B(\alpha)$ (Dicker, 2014a).

Recall that the ridge estimator with optimally tuned regularization parameter is

$$\hat{\beta}_r(\alpha) = (X^\top X + \frac{p}{\alpha^2} I_p)^{-1} X^\top Y,$$

which can be interpreted as the posterior mean of $\beta$ under the normal prior assumption $\beta \sim \mathcal{N}(0, \alpha^2/p I_p)$. When $p$ is very large, the normal distribution $\mathcal{N}(0, \alpha^2/p I_p)$ is very close to the uniform distribution on $\mathbb{S}^{p-1}(\alpha)$, so we would expect that $\hat{\beta}_{\mathbb{S}^{p-1}(\alpha)} \approx \hat{\beta}_r(\alpha)$. With this intuition, Dicker (2014a) further showed that, as $p, n \to \infty, p/n \to \gamma \in (0, \infty)$, for any $\beta \in \mathbb{S}^{p-1}(\alpha)$

$$\lim_{n,p \to \infty} \left[ R(\hat{\beta}_{\mathbb{S}^{p-1}(\alpha)}, \beta) - R(\hat{\beta}_r(\alpha), \beta) \right] = 0.$$

So the global ridge estimator is asymptotically exact minimax.

We call an estimator is *asymptotically rate-minimax* if asymptotically its risk is at most a constant times the minimax risk. For our distributed problem, we have the following result:

**Theorem 3.4.7** (Minimax optimality)**.** *For fixed signal strength $\alpha^2$, the optimally weighted distributed ridge estimator is asymptotically rate minimax. Specifically, its risk $\mathcal{M}_k$ is less than the risk $\mathcal{M}_1$ of the global ridge estimator multiplied by a constant $C = 1 + \alpha^2$ which only depends on the signal strength $\alpha^2$, and not on the aspect ratio $\gamma = \lim p/n$ and number of machines $k$. Specifically*

$$\mathcal{M}_k \leq (1 + \alpha^2)\mathcal{M}_1.$$

*Moreover, for fixed aspect ratio $\gamma > 1$, the distributed risk $\mathcal{M}_k$ is less than the global risk $\mathcal{M}_1$ times a constant $C' = \gamma^2/(\gamma^2 - 1)$ which is independent of $\alpha^2$ and $k$, i.e.*

$$\mathcal{M}_k \leq \frac{\gamma^2}{\gamma^2 - 1}\mathcal{M}_1.$$

*Therefore, in either case, the optimally weighted distributed ridge estimator is asymptotically rate minimax.*

See Section 3.7.15 for the proof. The minimax optimality result is nontrivial, and does not hold for some simpler estimators. For instance, for the null estimator $\hat{\beta}_{null} = 0$, the corresponding ARE can be written in terms of the optimal risk function $\phi(\gamma)$ as

$$\lim_{n,p\to\infty} \frac{R(\hat{\beta}_r(\alpha), \beta)}{R(\hat{\beta}_{null}, \beta)} = \frac{\phi(\gamma)}{\alpha^2} = \frac{\gamma m_\gamma(-\gamma/\alpha^2)}{\alpha^2}.$$

When $\gamma \to \infty$, we know that $\gamma/\alpha^2 m_\gamma(-\gamma/\alpha^2) \to 1$, so that even *the null estimator is asymptotically exact minimax.* In this regime, exact minimaxity is a weak result. When $\gamma \to 0$ however, we have $\gamma/\alpha^2 m_\gamma(-\gamma/\alpha^2) \to 0$ for any $\alpha$, and so the null estimator does not perform well (has zero efficiency). However, the distributed estimator is still asymptotically rate-minimax.

**Algorithm 1:** WONDER: Weighted ONe-shot DistributEd Ridge regression algorithm, general design

**Input :** Data matrices $(n_i \times p)$ and outcomes $(n_i \times 1)$, $(X_i, Y_i)$ distributed across $k$ sites

**Output:** Distributed ridge estimator $\hat{\beta}_{dist}$ of regression coefficients $\beta$

**for** $i \leftarrow 1$ **to** $k$ **do**
  Compute the MLE $\hat{\theta}_i = (\hat{\sigma}_i^2, \hat{\alpha}_i^2)$ locally on $i$-th machine;
  Send $\hat{\theta}_i$ to the global data center;
**end**
At the data center, combine $\hat{\theta}_i$ to get a global estimator $\hat{\theta} = (\hat{\sigma}^2, \hat{\alpha}^2) = k^{-1} \sum_{i=1}^k \hat{\theta}_i$ and send it back to the local machines;
Choose a set of tuning parameters $\mathcal{S}$ around the initial guess $\lambda_0 = kp/(n\hat{\alpha}^2)$;
**for** $\lambda \in \mathcal{S}$ **do**
  **for** $i \leftarrow 1$ **to** $k$ **do**
    Compute the local ridge estimator $\hat{\beta}_i(\lambda) = (X_i^\top X_i + n_i \lambda I_p)^{-1} X_i^\top Y_i$;
    Compute the weight $\omega_i$ for the $i$-th local estimator by using the formulas from Theorem 3.3.2:

$$\omega_i(\lambda) = \frac{\hat{\sigma}^2 \hat{\alpha}^2 (1 - \lambda m)}{\mathcal{F} + k\mathcal{G}}$$

    where we use $\mathrm{tr}[(X_i^\top X_i/n_i + \lambda I)^{-1}]/p$ to approximate $m$, and use $\mathrm{tr}[(X_i^\top X_i/n_i + \lambda I)^{-2}]/p$ to approximate $m'$;
    Send $\hat{\beta}_i(\lambda)$ and $\omega_i(\lambda)$ to the global data center;
  **end**
  Evaluate the performance of the distributed ridge estimator $\hat{\beta}_{dist}(\lambda) = \sum_{i=1}^k \omega_i(\lambda) \hat{\beta}_i(\lambda)$ on validation sets;
**end**
Select the best tuning parameter $\lambda^*$ and output the corresponding distributed ridge estimator $\hat{\beta}_{dist}(\lambda^*) = \sum_{i=1}^k \omega_i(\lambda^*) \hat{\beta}_i(\lambda^*)$.

So far, most of our results on distributed ridge regression are purely theoretical. In practice, it would be very helpful to have an implementable algorithm. In fact, our theory for distributed ridge regression allows us to develop an efficient algorithm which works for designs $X$ with arbitrary covariance structures $\Sigma$.

Recall that we have $n$ samples distributed across $k$ machines. For simplicity, let us assume the samples are equally distributed. On the $i$-th machine, we compute a local ridge estimator $\hat{\beta}_i$, local estimators $\hat{\sigma}_i^2$, $\hat{\alpha}_i^2$ of the signal-to-noise ratio and the noise level. From Theorem

3.3.2, we know that the other quantities needed to find the optimal weights are $m, m'$ and $\lambda$. For $m$ and $m'$, by the definition of the Stieltjes transform, they can be approximated by

$$\frac{\mathrm{tr}[(\widehat{\Sigma}_i + \lambda I)^{-1}]}{p} \approx m(-\lambda) \quad \text{and} \quad \frac{\mathrm{tr}[(\widehat{\Sigma}_i + \lambda I)^{-2}]}{p} \approx m'(-\lambda).$$

Here we only need to use local data. The remaining question is: how do we choose the tuning parameter $\lambda$? One way may be grid search. From the theory for the isotropic design, a proper initial guess would be $\lambda = kp/(n\alpha^2)$. Then we can search around this initial guess to find a good parameter with small prediction error.

We assume the data are already mean-centered, which can be performed exactly in one additional round of communication, or approximately by centering the individual data sets.

Now we have all the quantities we need for our Weighted ONe-shot DistributEd Ridge regression algorithm (WONDER). We send them to a global machine or data center, and aggregate them to compute a weighted ridge estimator. See Algorithm 1 for more details. WONDER is communication efficient as the local machines only need to send the local ridge estimator $\hat{\beta}_i$ and some scalars to the global datacenter.

For identity covariance, our results lead to a much simpler WONDER algorithm which requires even less communication and computation. See Algorithm 2.

In the above WONDER algorithms, we combine the local estimators of the noise level and signal strength $\hat{\theta}_i$ to find a global estimator $\hat{\theta}$. A simple method is to take the average: $\hat{\theta} = k^{-1} \sum_{i=1}^{k} \hat{\theta}_i$. Another option is to use inverse-variance weighting, based on the asymptotic variance of the MLE (which then of course has to be estimated).

Based on the results so far, it follows that our WONDER algorithm can consistently estimate the limiting optimal weights, and moreover it has asymptotically optimal mean squared error among all weighted distributed ridge estimators, at least for the identity covariance case. We omit the details.

**Algorithm 2:** WONDER: Weighted ONe-shot DistributEd Ridge regression algorithm, isotropic design

**Input :** Data matrices $(n_i \times p)$ and outcomes $(n_i \times 1)$, $(X_i, Y_i)$ distributed across $k$ sites

**Output:** Distributed ridge estimator $\hat{\beta}_{dist}$ of regression coefficients $\beta$

**for** $i \leftarrow 1$ **to** $k$ **do**
  Compute the MLE $\hat{\theta}_i = (\hat{\sigma}_i^2, \hat{\alpha}_i^2)$ locally on $i$-th machine;
  Set local aspect ratio $\gamma_i = p/n_i$;
  Set regularization parameter $\lambda_i = \gamma_i/\hat{\alpha}_i^2$;
  Compute the local ridge estimator $\hat{\beta}_i(\lambda_i) = (X_i^\top X_i + n_i\lambda_i I_p)^{-1} X_i^\top Y_i$;
  Send $\hat{\theta}_i, \gamma_i$ and $\hat{\beta}_i$ to the global data center.
**end**
At the data center, combine $\hat{\theta}_i$ to get a global estimator $\hat{\theta} = (\hat{\sigma}^2, \hat{\alpha}^2)$, by $\hat{\theta} = k^{-1}\sum_{i=1}^k \hat{\theta}_i$;
Evaluate the optimal risk functions for $i = 1, 2, \ldots, k$

$$\phi(\gamma_i) = \gamma_i m_{\gamma_i}(-\gamma_i/\hat{\alpha}^2) = \frac{-\gamma_i/\hat{\alpha}^2 + \gamma_i - 1 + \sqrt{(-\gamma_i/\hat{\alpha}^2 + \gamma_i - 1)^2 + 4\gamma_i^2/\hat{\alpha}^2}}{2\gamma_i/\hat{\alpha}^2};$$

Compute the optimal weights $\omega$, where the $i$-th coordinate of $\omega$ is

$$\omega_i = \left(\frac{\hat{\alpha}^2}{\phi(\gamma_i)}\right) \cdot \left(\frac{1}{1 + \sum_{i=1}^k \left[\frac{\hat{\alpha}^2}{\phi(\gamma_i)} - 1\right]}\right);$$

Output the distributed ridge estimator $\hat{\beta}_{dist} = \sum_{i=1}^k \omega_i \hat{\beta}_i$.

## 3.6. Experimental Results

We present some numerical results in addition to the ones already shown in the paper.

*3.6.1. Finite-sample comparison of relative efficiency for isotropic covariance*

Figure 15 shows a comparison of the theoretical formulas for ARE and realized relative efficiency in a regression simulation. Here the regression model is $Y = X\beta + \varepsilon$, where $X$ is $n \times p$ with i.i.d. standard normal entries, $\beta$ is a $p$-dimensional random vector with i.i.d. mean 0, variance $\alpha^2/p$ normal entries, and $\varepsilon$ also has i.i.d standard normal entries. For each $k = 1, 2, 5, 10, 20, 50$, we split the data equally into $k$ groups and perform ridge regression on each group. For each group, we choose the same tuning parameter $\lambda_i = p/(n_i\alpha^2)$. For the global regression on the entire data set, we choose the tuning parameter $\lambda = p/(n\alpha^2)$

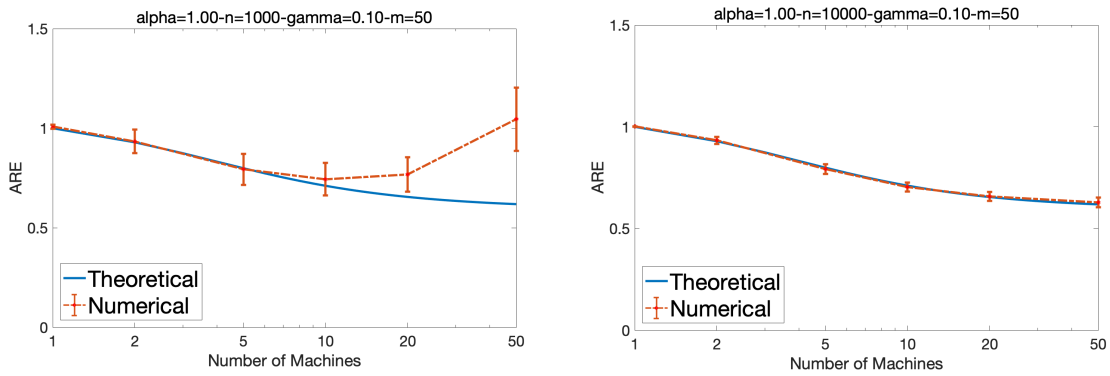



Figure 15: Realized relative efficiency in a regression simulation.

optimally.

We show the results of the expression for the realized relative efficiency $\|\hat{\beta} - \beta\|^2 / \|\hat{\beta}_{dist} - \beta\|^2$ compared to the theoretical ARE. We generate 100 independent copies of $\varepsilon$, perform regression, recording the realized relative efficiency $||\hat{\beta} - \beta||^2 / ||\hat{\beta}_{dist} - \beta||^2$, as well as its overall Monte Carlo mean. For the first plot, we take $n = 1000$, $p = 100$, and $\alpha = \sigma = 1$.

As we can see in the plot, the theoretical formula is accurate only for a small number of machines. It turns out that this is due to finite-sample effects. In the second plot, we set $n = 10000$, $p = 1000$ and $\alpha = \sigma = 1$ such that the aspect ratio $\gamma = p/n$ is the same as before. In that case the theoretical formula becomes very accurate.

*3.6.2. Choosing the regularization for general covariance*

How can we choose the optimal regularization parameters when the predictors have a general covariance structure $\Sigma$? In this case, our theoretical results do not give an explicit expression for the optimal regularization parameters. In practice, one can use techniques like cross-validation to do selections. Here we present simulation results to shed light on the important question of how to choose them.

We use a similar simulation setup as in the previous sections, except we generate the datapoints independently from an autoregressive model of order one (AR-1), i.e., each datapoint $x_i$ is generated as $x_i \sim \mathcal{N}(0, \Sigma)$, where $\Sigma_{ij} = \rho^{|i-j|}$, and $\rho$ is the autocorrelation

parameter. We choose $\rho = 0.9$. We also choose $n = 3000$, $p = 500$, and report the results of a simulation where we average over $n_{mc} = 20$ independent realizations of $\beta$. Figure 16 shows the optimal distributed risk $M^*(k)$ as a function of the local regularization parameter $\lambda$. We set all local regularization parameters to equal values, which is reasonable, since the local problems are exchangeable. We also parametrize the regularization parameters as multiples of the optimal parameter for the isotropic case (which equals $k\gamma/\alpha^2$).

We observe that for $k = 1$, the optimal parameter is the same as in the isotropic case. This makes sense, because the optimal regularization parameter for one machine is always the same, regardless of the structure of the design. However for $k > 1$, we observe that the regularization parameters are *smaller* than the isotropic ones. This is an insight that has apparently not been available before. It is an interesting topic of future work to develop an intuitive understanding.

*3.6.3. Experiments on empirical data*

In this section, we present an empirical data example to examine the accuracy of our theoretical results. It is reasonable to compare the performance of different estimators in terms of the prediction (test) error. Figure 17 shows a comparison of three estimators including our optimal weighted estimator on the Million Song Year Prediction Dataset (MSD) (Bertin-Mahieux et al., 2011).

Specifically, we perform the following steps in our data analysis. We download the data set from the UC Irvine Machine Learning Repository. The original data set has $N = 515,345$ samples and $p = 91$ features. The data set has already been divided into a training set and a test set. The training set consists of the first $463,715$ samples and the test set contains the rest. We attempt to predict the release year of a song. Before doing distributed regression, we first center and normalize both the design matrix $X$ and the outcome $Y$. Now we are ready to do ridge regression under the distributed setting.

For each experiment, we randomly choose $n_{train} = 10,000$ samples from the training set and $n_{test} = 1,000$ samples from the test set. We construct the estimators on the training

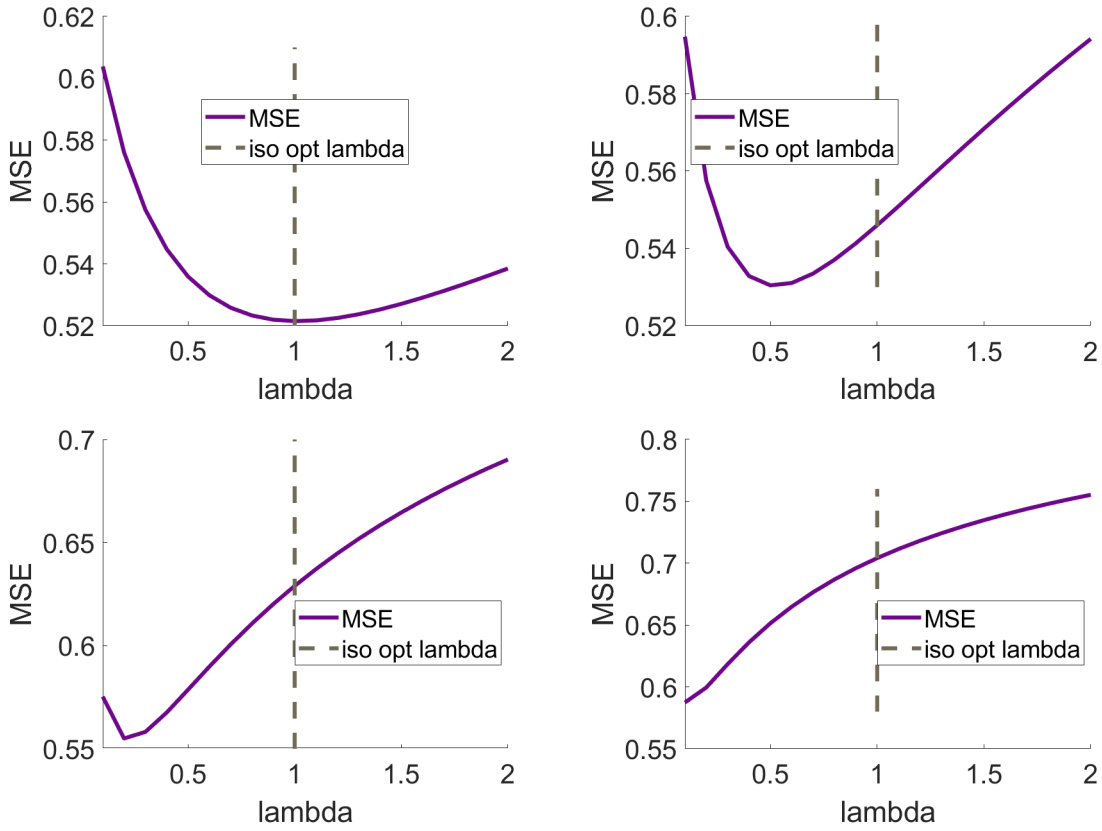



Figure 16: Distributed risk as a function of the regularization parameter. We plot the risk of the optimally weighted distributed estimator for an AR-1 covariance structure. We set $\alpha = 1$, $\gamma = 0.17$ and $k = 1, 2, 5, 10$.

samples. Then we perform ridge regression in a distributed way to obtain our optimal weighted WONDER estimator as described in Algorithm 1. We measure its performance on the test data by computing its MSE for prediction. We choose the number of machines to be $k = 1, 2, 5, 10, 20, 50, 100, 200, 500, 1000$, and we distribute the data evenly across the $k$ machines. Here we try different tuning parameters $\lambda$ around $kp/(n_{train} \cdot \hat{\alpha}^2)$, and use $\lambda = 3kp/(n_{train} \cdot \hat{\alpha}^2)$ as our final parameter. (In practice, one may try a 1-D grid search to find the right scale.)

For comparison, we also consider two other estimators:

1. The distributed estimator where we take the naive average (weight for each local estimator is simply $1/k$) and choose the local tuning parameter $\lambda = p/(n_{train} \cdot \hat{\alpha}^2)$. This formally

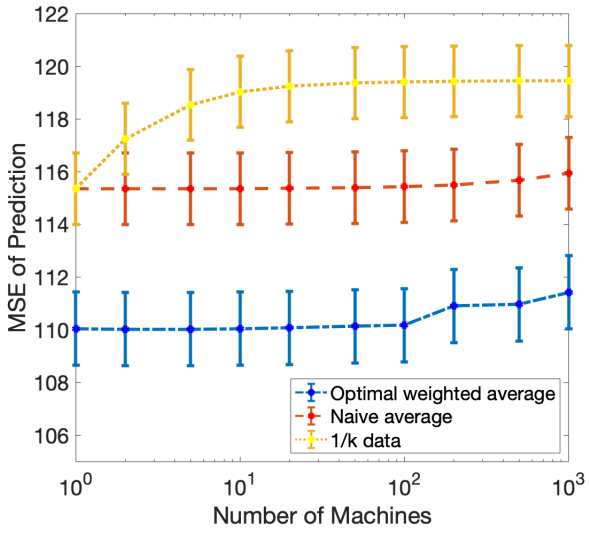



Figure 17: Million Song Year Prediction Dataset (MSD). Optimal weighted average (WONDER), Naive average, and regression on $1/k$ fraction of data.

agrees with the divide-and-conquer type estimator proposed in Zhang et al. (2015).

2. The estimator using only a fraction $1/k$ of the data, which is just one of the local estimators. For this estimator, we choose the tuning parameter $\lambda = kp/(n_{train} \cdot \hat{\alpha}^2)$.

We repeat the experiment for $T = 100$ times, and report the average and $1/4$ standard deviation over all experiments on Figure 17. Each time we randomly collect new training and test sets.

From Figure 17, we observe the following:

1. The WONDER estimator has smaller MSE than both the local estimator and the naive averaged estimator, which means optimal weighting can indeed help.

2. It seems that data splitting does not have huge impact on the performance of the WONDER estimator. This phenomenon is compatible with our theory. Since the signal-to-noise ratio $\alpha^2$ is about 1.2 for this data set, we are in a low SNR scenario. From Proposition 3.4.6 and Figure 14, we see that the performance of the distributed estimator is close to the global estimator in terms of the prediction error.

To conclude, in terms of computation-statistics tradeoff, this example suggests a very positive outlook on using distributed ridge regression via WONDER: The accuracy is affected very little even though the data is split up into 100 parts. Thus we save at least 100x in computation time, while we have nearly no loss in performance.

Finally, we mention that in Figure 4 of Zhang et al. (2015), the authors also compare the performance of the distributed estimator to the local estimator on the same Million Song data set. We notice that the MSE of prediction in their experiments is usually between 80 and 90, and variance is typically very small. In our experiments, both the MSE and variance are larger. The reason for this seems to be that they consider more general kernel ridge regression.

## 3.7. Appendix

### *3.7.1. Adding a Constant to the Regression*

We show below the details of the derivation of optimal weights for ridge regression when we also add a constant to the (biased) local estimators. In our calculation from Theorem 3.2.1, we need to change some details as follows:

We need to define a new matrix $\widehat{B} = [\hat{\beta}_1, \ldots, \hat{\beta}_k, p^{-1/2}1_p]$ and new weights $w = [w; w_{k+1}]$. Clearly, we still have that

$$B = [\mathbb{E}\hat{\beta}_1, \ldots, \mathbb{E}\hat{\beta}_k, p^{-1/2}1_p] = [Q_1\beta; \ldots; Q_k\beta, p^{-1/2}1_p].$$

The new matrix $R$ is now diagonal with all entries as before, and the lower right corner entry is $R_{k+1} = 0$.

We consider the same regression problem as before, except we add an intercept into the matrix $B$ as above. The same algebraic form of the optimal weights and risk holds, with the new definitions above. The optimal risk is now

$$M^*(k) = \|\beta\|^2 - v^\top (A+R)^{-1} v$$

where

$$\begin{aligned}
v &= B^\top \beta = [vec[\beta^\top Q_i \beta]; p^{-1/2} 1_p^\top \beta] \\
A &= \begin{bmatrix} mx[\beta^\top Q_i Q_j \beta] & vec[p^{-1/2} 1_p^\top Q_i \beta] \\ vec[p^{-1/2} 1_p^\top Q_i \beta] & 1 \end{bmatrix} \\
R &= \operatorname{diag}\left[ n_i^{-1} \operatorname{tr}[(\widehat{\Sigma}_i + \lambda_i I_p)^{-2} \widehat{\Sigma}_i]; 0 \right] \\
Q_i &= (\widehat{\Sigma}_i + \lambda_i I_p)^{-1} \widehat{\Sigma}_i
\end{aligned}$$

In simulation studies, we have observed that this approach typically does not lead to a significant decrease in MSE.

*3.7.2. Differentiation Rule for Calculus of Deterministic Equivalents*

**Theorem 3.7.1** (Differentiation rule)**.** *Suppose $T = T_n$ and $S = S_n$ are two (deterministic or random) matrix sequences of growing dimensions such that $f(z, T_n) \asymp g(z, S_n)$, where the entries of $f$ and $g$ are analytic functions in $z \in D$ and $D$ is an open connected subset of $\mathbb{C}$. Suppose that for any sequence $C_n$ of deterministic matrices with bounded trace norm we have*

$$|\operatorname{tr}[C_n(f(z, T_n) - g(z, S_n))]| \leq M$$

*for every $n$ and $z \in D$. Then we have $f'(z, T_n) \asymp g'(z, S_n)$ for $z \in D$, where the derivatives are entry-wise with respect to $z$.*

To prove this theorem, we need to introduce a lemma from complex analysis which is a consequence of the dominated convergence theorem and Cauchy's integral formula.

**Lemma 3.7.2** (see Lemma 2.14 in Bai and Silverstein (2010))**.** *Let $f_1, f_2, \ldots$ be analytic on the domain $D$, satisfying $|f_n(z)| \leq M$ for every $n$ and $z \in D$. Suppose that there is an analytic function on $D$ such that $f_n(z) \to f(z)$ for all $z \in D$. Then it also holds that*

$f_n'(z) \to f'(z)$ *for all* $z \in D$.

The proof of theorem 3.7.1 is clear. Since $\operatorname{tr}[C_n(f(z,T_n) - g(z,S_n))]$ is a sequence of analytic functions on $D$ with uniform bound, then from the definition of the deterministic equivalence, we have $\operatorname{tr}[C_n(f(z,T_n) - g(z,S_n))] \to 0$. By lemma 3.7.2, the derivative also converges to 0 for all $z \in D$, which finishes the proof.

*3.7.3. Gaussian MLE for Signal and Noise Components*

Recall that our model is $Y = X\beta + \varepsilon$ where $\beta$ and $\varepsilon$ are independent. Let $\theta = (\sigma^2, \alpha^2)$ and define the Gaussian log-likelihood,

$$\ell(\theta) = -\frac{1}{2}\log(\sigma^2) - \frac{1}{2n}\log\det\left(\frac{\alpha^2}{p}XX^\top + I\right) - \frac{1}{2\sigma^2 n}Y^\top\left(\frac{\alpha^2}{p}XX^\top + I\right)^{-1}Y.$$

Note that $\ell(\theta)$ is the log-likelihood for $\theta$ under the Gaussian assumption of $\beta \sim \mathcal{N}(0, (\sigma^2\alpha^2/p)I)$ and $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$. For the MLE

$$\hat{\theta} = (\hat{\sigma}^2, \hat{\alpha}^2) = \underset{\sigma^2,\alpha^2 \geq 0}{\operatorname{argmax}}\ \ell(\theta),$$

we have the following result from Dicker and Erdogdu (2017).

**Theorem 3.7.3** (Consistency and asymptotic normality, Dicker and Erdogdu (2017))**.** *Suppose* $\theta = (\sigma^2, \alpha^2)$ *are the true parameters, then* $\hat{\theta} \to \theta$ *in probability as* $p/n \to \gamma$. *Furthermore, define the Fisher information matrix for* $\theta$ *under the Gaussian assumption model*

$$\mathcal{I}_n(\theta) = \begin{bmatrix} I_2(\theta) & I_3(\theta) \\ I_3(\theta) & I_4(\theta) \end{bmatrix},$$

*where*

$$I_k(\theta) = \frac{1}{2n\sigma^{8-2k}}\operatorname{tr}\left[\left(\frac{1}{p}XX^\top\right)^{k-2}\left(\frac{\alpha^2}{p}XX^\top + I\right)^{2-k}\right], \quad k = 2,3,4.$$

*Then* $n^{1/2}\mathcal{I}_n(\theta)^{1/2}(\hat{\theta} - \theta) \to \mathcal{N}(0, I_2)$ *in distribution as* $p/n \to \gamma$.

In addition, if we put some assumptions on $X$ as we did in Theorem 2 and denote the

limiting spectral distribution of $p^{-1}XX^\top$ by $F_\gamma$, then the entries of the Fisher information matrix $\mathcal{I}_n(\theta)$ have limits

$$I_k(\theta) \to_{a.s.} \mathcal{J}_k(\theta) = \frac{1}{2\sigma^{8-2k}} \int \left(\frac{s}{\alpha^2 s + 1}\right)^{k-2} dF_\gamma(s), \quad k = 2, 3, 4.$$

Thus $\mathcal{I}_n(\theta)$ converges almost surely to a limiting information matrix $\mathcal{I}(\theta)$ which characterizes the asymptotic variance of the MLE $\hat{\theta}$.

*3.7.4. Explaining Decoupling via Free Probability Theory*

In this section, we provide an explanation via free probability theory for why the limiting distributed risk decouples. Specifically, we explain why the limit of the quantities $\beta^\top Q_i \beta \cdot \beta^\top Q_j \beta$ for $i \neq j$ becomes a product of terms depending on $i, j$.

We will use some basic notions from free probability theory (Voiculescu et al., 1992; Hiai and Petz, 2006; Nica and Speicher, 2006; Anderson et al., 2009; Couillet and Debbah, 2011). Let us define our non-commutative probability space as

$$\left(\mathcal{A} = (L^{\infty-} \otimes M_p(\mathbb{R})), \tau = \frac{1}{p}\operatorname{tr}\right),$$

where $L^{\infty-}$ denotes the collection of random variables with all moments finite and $M_p(\mathbb{R})$ is the space of $p \times p$ real matrices. Recall that, a sequence of random variables $\{a_{1,p}, a_{2,p}, \ldots, a_{k,p}\} \subset \mathcal{A}$ is said to be asymptotically free almost surely if

$$\tau[\prod_{j=1}^{m} P_j(a_{i_j,p} - \tau(P_j(a_{i_j,p})))] \to_{a.s.} 0,$$

for any positive integer $m$, any polynomials $P_1, \ldots, P_m$ and any indices $i_1, \ldots, i_m \in [k]$ with no two adjacent $i_j$ equal. Suppose $A_p, B_p$ are two sequences of independent random matrices and at least one of them is orthogonally invariant, then it is well-known that $\{A_p, B_p\} \subset \mathcal{A}$ is asymptotically free almost surely.

Now, let us assume that $X^\top X$ is orthogonally invariant, which is the case when $X^\top X$

follows the white Wishart distribution. Then clearly $X_i^\top X_i$ and $X_j^\top X_j$ are asymptotically free almost surely. It follows that $Q_i$ and $Q_j$ are also asymptotically free almost surely. By using the definition of asymptotic freeness, we have for $i \neq j$

$$\tau[(Q_i - \frac{1}{p}\operatorname{tr}(Q_i)I)(Q_j - \frac{1}{p}\operatorname{tr}(Q_j)I)] \to_{a.s.} 0,$$

which is equivalent to

$$\frac{1}{p}\operatorname{tr}(Q_iQ_j) - \frac{1}{p}\operatorname{tr}(Q_i)\frac{1}{p}\operatorname{tr}(Q_j) \to_{a.s.} 0.$$

Hence, under the random-effects assumption for $\beta$, the limit of $\beta^\top\beta \cdot \beta^\top Q_iQ_j\beta$ $(i \neq j)$ will decouple and is the same as the limit of $\beta^\top Q_i\beta \cdot \beta^\top Q_j\beta$.

*3.7.5. Intuitive Explanation for the Need to Use Weights Summing to Greater than Unity*

Consider a much more simplified problem, where we are estimating a scalar parameter $\theta$. We have an estimator $\hat{\theta}$, which is generally biased, and we would like to find the scale multiple $c \cdot \hat{\theta}$ that minimizes the mean squared error. A calculation reveals that

$$M(c) = \mathbb{E}(c \cdot \hat{\theta} - \theta)^2 = c^2\mathbb{E}(\hat{\theta}^2) - 2c \cdot \mathbb{E}\hat{\theta} \cdot \theta + \theta^2$$

Hence the optimal scale factor is $c = \mathbb{E}\hat{\theta} \cdot \theta / \mathbb{E}(\hat{\theta}^2)$.

We can achieve a better understanding of this optimal scale if we consider the bias-variance decomposition of $\hat{\theta}$. Let us define the bias and the variance as

$$B = \mathbb{E}\hat{\theta} - \theta$$
$$V = \mathbb{E}(\hat{\theta} - \mathbb{E}\hat{\theta})^2$$

We then see that the optimal scale factor is

$$c = \frac{B+\theta}{V+(B+\theta)^2}\theta = 1 - \frac{V+B(B+\theta)}{V+(B+\theta)^2}.$$

This quantity is an "inflation factor", i.e., greater than or equal to unity, if $V+B(B+\theta) \leq 0$. This can be written as

$$V + B^2 \leq -B\theta$$

Hence, this condition can only hold if the bias $B$ has opposite sign with $\theta$. This would be the case for a *shrinkage estimator* $\theta$. In that case, the condition could hold if the parameter $\theta$ has a large magnitude.

Returning to our main problem, ridge regression is a shrinkage estimator, and averages of ridge regression estimators are still shrinkage estimators. Therefore, it makes sense that their weighted average should be inflated to minimize mean squared error. This provides an intuitive explanation for why the weights sum to greater than one.

*3.7.6. Proof of Theorem 3.2.1*

We can calculate the MSE of the weighted sum as

$$\begin{aligned} M(w) = \mathbb{E}\left\|\sum w_i\hat{\beta}_i - \beta\right\|^2 &= \mathbb{E}(\sum w_i\hat{\beta}_i - \beta)^\top(\sum w_j\hat{\beta}_j - \beta) \\ &= \sum_{ij} w_i w_j \cdot \mathbb{E}\hat{\beta}_i^\top\hat{\beta}_j - 2\sum_i w_i\mathbb{E}\hat{\beta}_i^\top\beta + \|\beta\|^2. \end{aligned}$$

Let $\widehat{B}$ be the $p \times k$ matrix defined as $\widehat{B} = [\hat{\beta}_1, \ldots, \hat{\beta}_k]$. Then we can write the above MSE as

$$M(w) = w^\top\mathbb{E}\widehat{B}^\top\widehat{B}w - 2\mathbb{E}\beta^\top\widehat{B}w + \|\beta\|^2.$$

Let also

$$B = \mathbb{E}\widehat{B} = [\mathbb{E}\hat{\beta}_1, \ldots, \mathbb{E}\hat{\beta}_k].$$

Since the local estimators are independent, we can write

$$M(w) = w^\top (B^\top B + R)w - 2\beta^\top Bw + \|\beta\|^2,$$

where $R$ is a diagonal matrix with entries

$$R_i = \mathbb{E}\|\hat{\beta}_i\|^2 - \|\mathbb{E}\hat{\beta}_i\|^2 = \mathbb{E}\|\hat{\beta}_i - \mathbb{E}\hat{\beta}_i\|^2.$$

The objective function $M(w)$ can be viewed as corresponding to a $k$-parameter linear regression problem, with unknown parameters $w_i$, design matrix $B$ and outcome vector $\beta$. Specifically, we regress $\beta$ on $\mathbb{E}\widehat{B} = \mathbb{E}[\hat{\beta}_1, \ldots, \hat{\beta}_k]$. Therefore, the optimal weights are

$$w^* = (B^\top B + R)^{-1} B^\top \beta,$$

and the optimal risk equals

$$M^* = M(w^*) = \beta^\top \left[ I_p - B(B^\top B + R)^{-1} B^\top \right] \beta.$$

Now, to find $B = \mathbb{E}\widehat{B}$, we need $\mathbb{E}\hat{\beta}_i$. The expectation of the ridge regression estimator for the full data set is

$$\mathbb{E}\hat{\beta}(\lambda) = \mathbb{E}(X^\top X + n\lambda I_p)^{-1} X^\top Y = (X^\top X + n\lambda I_p)^{-1} X^\top X \beta.$$

Letting $\widehat{\Sigma} = n^{-1} X^\top X$, this equals $\mathbb{E}\hat{\beta}(\lambda) = (\widehat{\Sigma} + \lambda I_p)^{-1} \widehat{\Sigma}\beta$. Similarly,

$$\mathbb{E}\hat{\beta}_i(\lambda_i) = (X_i^\top X_i + n_i \lambda_i I_p)^{-1} X_i^\top X_i \beta.$$

Let $Q_i = Q_i(\lambda_i) = (X_i^\top X_i + n_i \lambda_i I_p)^{-1} X_i^\top X_i$ be those matrices and let $\widehat{\Sigma}_i = n^{-1} X_i^\top X_i$. Then the above equals $Q_i = (\widehat{\Sigma}_i + \lambda_i I_p)^{-1} \widehat{\Sigma}_i$, and

$$B = [Q_1\beta; \ldots; Q_k\beta].$$

Therefore, $B^\top B$ has entries $\beta^\top Q_i Q_j \beta$, while $B^\top \beta$ has entries $\beta^\top Q_i \beta$. Moreover,

$$R_i = \mathbb{E}\|\hat{\beta}_i - \mathbb{E}\hat{\beta}_i\|^2 = \mathbb{E}\|(X_i^\top X_i + n_i\lambda_i I_p)^{-1} X_i^\top \varepsilon_i\|^2 = \sigma^2 \operatorname{tr}[(X_i^\top X_i + n_i\lambda_i I_p)^{-2} X_i^\top X_i].$$

We can also write this as $R_i = n_i^{-1}\sigma^2 \operatorname{tr}[(\widehat{\Sigma}_i + \lambda_i I_p)^{-2}\widehat{\Sigma}_i]$. To conclude the optimal risk, we have

$$M^*(k) = \|\beta\|^2 - v^\top (A+R)^{-1} v,$$

where

$$\begin{aligned} v &= B^\top \beta = vec[\beta^\top Q_i \beta], \\ A &= mat[\beta^\top Q_i Q_j \beta], \\ R &= \operatorname{diag}\left[n_i^{-1}\sigma^2 \operatorname{tr}[(\widehat{\Sigma}_i + \lambda_i I_p)^{-2}\widehat{\Sigma}_i]\right], \\ Q_i &= (\widehat{\Sigma}_i + \lambda_i I_p)^{-1}\widehat{\Sigma}_i. \end{aligned}$$

Here we used the vectorization and to-matrix operators $vec, mat$. For the global MSE, we only need to consider the special case where $k = 1$, which gives us

$$\mathbb{E}||\hat{\beta} - \beta||^2 = M^*(1) = \|\beta\|^2 - \frac{(\beta^\top Q \beta)^2}{\beta^\top Q^2 \beta + \sigma^2 \operatorname{tr}[(X^\top X + n\lambda I_p)^{-2} X^\top X]},$$

where $Q = (\widehat{\Sigma} + \lambda I_p)^{-1}\widehat{\Sigma}$. This finishes the argument.

*3.7.7. Proof of Theorem 3.3.1*

The first step is to use the well-known concentration of quadratic forms to reduce to trace functionals (See e.g. Lemma C.3 of Dobriban and Wager (2018) which is based on Lemma B.26 of Bai and Silverstein (2010)). Since $\beta$ is independent of the data $X$ with mean zero and finite variance, under the moment assumptions imposed in the theorem, we have

$$\beta^\top Q_i \beta - \sigma^2 \alpha^2 / p \cdot \operatorname{tr} Q_i \to_{a.s.} 0,$$

$$\beta^\top Q_i Q_j \beta - \sigma^2 \alpha^2 / p \cdot \operatorname{tr} Q_i Q_j \to_{a.s.} 0,$$

$$\beta^\top Q_i^2 \beta - \sigma^2\alpha^2/p \cdot \operatorname{tr} Q_i^2 \to_{a.s.} 0.$$

Let us compute the limits of $v, A$ and $R$ respectively.

1. Limit of $v$: First of all, we have already known that

$$\beta^\top Q_i \beta - \sigma^2\alpha^2/p \cdot \operatorname{tr} Q_i \to_{a.s.} 0,$$

so it is sufficient to consider the limit of $\operatorname{tr} Q_i/p$. Since

$$\operatorname{tr} Q_i/p = 1 - \lambda_i \operatorname{tr}[(\widehat{\Sigma}_i + \lambda_i I_p)^{-1}]/p.$$

assuming that the spectral distribution of $\widehat{\Sigma}_i$ converges almost surely to $F_{\gamma_i}$, we thus have

$$\operatorname{tr} Q_i/p \to_{a.s.} 1 - \lambda_i \mathbb{E}_{F_{\gamma_i}}(T + \lambda_i)^{-1} = 1 - \lambda_i m_{F_{\gamma_i}}(-\lambda_i).$$

Above we have introduced the Stieltjes transform $m_{F_{\gamma_i}}$ as a limiting object. So,

$$\beta^\top Q_i \beta \to_{a.s.} \sigma^2\alpha^2[1 - \lambda_i m_{F_{\gamma_i}}(-\lambda_i)].$$

For the form in terms of the population spectral distribution $H$, if $p/n \to \gamma$ and the spectral distribution of $\Sigma$ converges to $H$, we have by the general Marchenko-Pastur (MP) theorem of Rubio and Mestre (Rubio and Mestre, 2011), that

$$(\widehat{\Sigma} + \lambda I)^{-1} \asymp (x_p \Sigma + \lambda I)^{-1},$$

where $x_p$ is the unique positive solution of the fixed point equation

$$1 - x_p = \frac{x_p}{n} \operatorname{tr}\left[\Sigma(x_p\Sigma + \lambda I)^{-1}\right].$$

When $n, p \to \infty$, $x_p \to x$ and $x$ satisfies the equation

$$1 - x = \gamma\left[1 - \lambda \int_0^\infty \frac{dH(t)}{xt + \lambda}\right].$$

We remark that the assumptions made in the theorem suffice for using the Rubio-Mestre result. Moreover, we only use a special case of their result, similar to Dobriban and Sheng (2018). Hence from the calculus of deterministic equivalents (Dobriban and Sheng, 2018), we can take the traces of the matrices in question to obtain

$$\operatorname{tr} Q_i/p = 1 - \lambda_i \operatorname{tr}[(\widehat{\Sigma}_i + \lambda_i I_p)^{-1}]/p \asymp 1 - \lambda_i \operatorname{tr}[(x_i\Sigma + \lambda_i I)^{-1}]/p \to_{a.s.} \mathbb{E}_H \frac{x_i T}{x_i T + \lambda_i},$$

where $x_i = x(H, \gamma_i, -\lambda_i)$ is the unique solution of

$$1 - x_i = \gamma_i\left[1 - \lambda_i \int_0^\infty \frac{dH(t)}{x_i t + \lambda_i}\right].$$

2. Limit of $A$: Let us consider the cases $i \neq j$ and $i = j$ separately.

   (a) $i \neq j$: We begin by

   $$\beta^\top Q_i Q_j \beta - \sigma^2\alpha^2/p \cdot \operatorname{tr} Q_i Q_j \to_{a.s.} 0.$$

   Based on the above expression for $Q_i$, we have

   $$Q_i Q_j = I_p - \lambda_i(\widehat{\Sigma}_i + \lambda_i I_p)^{-1} - \lambda_j(\widehat{\Sigma}_j + \lambda_j I_p)^{-1} + \lambda_i\lambda_j(\widehat{\Sigma}_i + \lambda_i I_p)^{-1}(\widehat{\Sigma}_j + \lambda_j I_p)^{-1}.$$

So the key will be to find the limit of

$$E_{ij} = p^{-1} \operatorname{tr}\{(\widehat{\Sigma}_i + \lambda_i I_p)^{-1}(\widehat{\Sigma}_j + \lambda_j I_p)^{-1}\}.$$

From the general MP theorem, since $p/n_i \to \gamma_i$, we have for all $i$,

$$(\widehat{\Sigma}_i + \lambda_i I_p)^{-1} \asymp (x_{ip}\Sigma + \lambda_i I_p)^{-1}.$$

Here $x_{ip}$ is the unique positive solution of the fixed point equation

$$1 - x_{ip} = \frac{x_{ip}}{n_i} \operatorname{tr}\left[\Sigma(x_{ip}\Sigma + \lambda_i I)^{-1}\right],$$

and $x_{ip} \to x_i$ as $n_i, p \to \infty$. By the product rule of the calculus of deterministic equivalents, we have for $i \neq j$

$$(\widehat{\Sigma}_i + \lambda_i I_p)^{-1}(\widehat{\Sigma}_j + \lambda_j I_p)^{-1} \asymp (x_{ip}\Sigma + \lambda_i I_p)^{-1}(x_{jp}\Sigma + \lambda_j I_p)^{-1}.$$

Hence by the trace rule of deterministic equivalents,

$$E_{ij} \asymp p^{-1} \operatorname{tr}[(x_{ip}\Sigma + \lambda_i I_p)^{-1}(x_{jp}\Sigma + \lambda_j I_p)^{-1}]$$

Moreover, since the spectral distribution of $\Sigma$ converges to $H$, we find for $i \neq j$

$$E_{ij} \to \mathbb{E}_H \frac{1}{(x_i T + \lambda_i)(x_j T + \lambda_j)}.$$

Putting it together,

$$Q_i Q_j \asymp I_p - \lambda_i(x_{ip}\Sigma + \lambda_i I_p)^{-1} - \lambda_j(x_{jp}\Sigma + \lambda_j I_p)^{-1} + \lambda_i\lambda_j(x_{ip}\Sigma + \lambda_i I_p)^{-1}(x_{jp}\Sigma + \lambda_j I_p)^{-1}.$$

So, again by the trace rule of deterministic equivalents, we have

$$p^{-1}\operatorname{tr}\{Q_iQ_j\} \to_{a.s.} 1 - \mathbb{E}_H\frac{\lambda_i}{x_iT+\lambda_i} - \mathbb{E}_H\frac{\lambda_j}{x_jT+\lambda_j} + \mathbb{E}_H\frac{\lambda_i\lambda_j}{(x_iT+\lambda_i)(x_jT+\lambda_j)}$$
$$= x_ix_j\mathbb{E}_H\frac{T^2}{(x_iT+\lambda_i)(x_jT+\lambda_j)}.$$

Therefore, for $i \neq j$

$$A_{ij} \to \sigma^2\alpha^2\left[x_ix_j\mathbb{E}_H\frac{T^2}{(x_iT+\lambda_i)(x_jT+\lambda_j)}\right].$$

(b) $i = j$: In this case,

$$\beta^\top Q_i^2\beta - \sigma^2\alpha^2/p \cdot \operatorname{tr} Q_i^2 \to 0,$$

where $Q_i^2 = I_p - 2\lambda_i(\widehat{\Sigma}_i + \lambda_iI_p)^{-1} + \lambda_i^2(\widehat{\Sigma}_i + \lambda_iI_p)^{-2}$. We can easily find the limit of $\operatorname{tr} Q_i^2/p$ in terms of empirical quantities, based on our knowledge of the convergence of Stieltjes transforms and its derivatives:

$$\operatorname{tr} Q_i^2/p \to 1 - 2\lambda_i m_{F_{\gamma_i}}(-\lambda_i) + \lambda_i^2 m'_{F_{\gamma_i}}(-\lambda_i).$$

Therefore, for $i = j$

$$A_{ii} \to \sigma^2\alpha^2[1 - 2\lambda_i m_{F_{\gamma_i}}(-\lambda_i) + \lambda_i^2 m'_{F_{\gamma_i}}(-\lambda_i)].$$

We can also express the limit of $A_{ii}$ in terms of the population spectral distribution $H$ by using Theorem 3.7.1. For our purpose, let $T = \Sigma$, $S = \widehat{\Sigma}$, while

$$f(z,T) = (x_pT - zI)^{-1},$$
$$g(z,S) = (S - zI_p)^{-1}.$$

From Rubio and Mestre (2011), we know that for each $z \in D := \mathbb{C} \setminus \mathbb{R}^+$, $f(z,\Sigma) \asymp$

$g(z, \widehat{\Sigma})$. $x_p$ is defined as

$$x_p = \frac{1}{n}\operatorname{tr}[(I + \frac{p}{n}e_p I)^{-1}] = \frac{1}{1 + (p/n)e_p} = \frac{1}{1 + \gamma_p e_p},$$

and $e_p = e_p(z)$ is the Stieltjes transform of a certain positive measure on $\mathbb{R}^+$, obtained as the unique solution of the equation

$$e_p = \frac{1}{p}\operatorname{tr}[\Sigma(x_p\Sigma - zI_p)^{-1}].$$

It is well-known that $x_p(z), e_p(z)$ are both analytic functions on $D$. Then we can check that the conditions of theorem 3.7.1 hold in this case. First of all, for an invertible matrix $A$, $A^{-1} = (\det A)^{-1}A^*$, where $A^*$ is the adjugate matrix of $A$. Since $x_p$ is analytic, it is easy to verify that $\det(x_p\Sigma - zI_p), \det(\widehat{\Sigma} - zI_p)$ and all entries of $(x_p\Sigma - zI_p)^*, (\widehat{\Sigma} - zI_p)^*$ are analytic functions of $z$. So the entries of $f(z, \Sigma)$ and $g(z, \widehat{\Sigma})$ are analytic in $D$.

Next, we want to bound

$$\operatorname{tr}[C_n((x_p\Sigma - zI_p)^{-1} - (\widehat{\Sigma} - zI_p)^{-1})] \leq ||C_n||_{\operatorname{tr}} \cdot ||(x_p\Sigma - zI_p)^{-1} - (\widehat{\Sigma} - zI_p)^{-1}||_2.$$

For a fixed $\delta > 0$, let us define a domain $D_\delta := \{z \in D : \operatorname{Re} z < -\delta\} \cup \{z \in D : |\operatorname{Im} z| > \delta\}$. Then, it is sufficient to find a uniform bound for $||(x_p\Sigma - zI_p)^{-1} - (\widehat{\Sigma} - zI_p)^{-1}||_2$ on $D_\delta$. In fact, we can bound $||(x_p\Sigma - zI_p)^{-1}||_2$ and $||(\widehat{\Sigma} - zI_p)^{-1}||_2$ separately.

i. Bounding $||(\widehat{\Sigma} - zI_p)^{-1}||_2$:

$$||(\widehat{\Sigma} - zI_p)^{-1}||_2 = \sigma_{\max}((\widehat{\Sigma} - zI_p)^{-1}) = \max_i \frac{1}{|\hat{l}_i - z|},$$

where $\hat{l}_i$ is the $i$-th eigenvalue of $\widehat{\Sigma}$. Since $\hat{l}_i$ is always non-negative, we have

$$\frac{1}{|\hat{l}_i - z|} = \frac{1}{|\hat{l}_i - \mathrm{Re}z - i\mathrm{Im}z|} = \frac{1}{\sqrt{(\hat{l}_i - \mathrm{Re}z)^2 + (\mathrm{Im}z)^2}} \leq \frac{1}{\delta}.$$

ii. Bounding $||(x_p\Sigma - zI_p)^{-1}||_2$:

In this case, we need to use the properties of $e_p$ and $x_p$. Recall that $e_p$ is the Stieltjes transform of a certain measure on $\mathbb{R}^+$, i.e.

$$\begin{aligned} e_p(z) &= \int_0^\infty \frac{1}{t-z} d\mu(t) = \int_0^\infty \frac{1}{t - \mathrm{Re}z - i\mathrm{Im}z} d\mu(t) \\ &= \int_0^\infty \frac{t - \mathrm{Re}z}{(t-\mathrm{Re}z)^2 + (\mathrm{Im}z)^2} d\mu(t) + i\int_0^\infty \frac{\mathrm{Im}z}{(t-\mathrm{Re}z)^2 + (\mathrm{Im}z)^2} d\mu(t). \end{aligned}$$

So

$$\begin{aligned} x_p &= \frac{1}{1+\gamma_p e_p} = \frac{1}{1 + \gamma_p \mathrm{Re}(e_p) + i\gamma_p \mathrm{Im}(e_p)} \\ &= \frac{1 + \gamma_p \mathrm{Re}(e_p)}{(1+\gamma_p \mathrm{Re}(e_p))^2 + (\gamma_p \mathrm{Im}(e_p))^2} - i\frac{\gamma_p \mathrm{Im}(e_p)}{(1+\gamma_p \mathrm{Re}(e_p))^2 + (\gamma_p \mathrm{Im}(e_p))^2}. \end{aligned}$$

When $z \in D_\delta$, we can check that $\mathrm{Re}(x_p) > 0$. Meanwhile, $\mathrm{Im}(x_p)$ and $\mathrm{Im}(z)$ have opposite signs.

Now, let us consider

$$||(x_p\Sigma - zI_p)^{-1}||_2 = \sigma_{\max}((x_p\Sigma - zI_p)^{-1}) = \max_k \frac{1}{|x_p l_k - z|},$$

where $l_k$ is the $k$-th eigenvalue of $\Sigma$. Since $l_k$ is non-negative, we have

$$\begin{aligned}\frac{1}{|x_p l_k - z|} &= \frac{1}{|l_k \mathrm{Re}(x_p) + i l_k \mathrm{Im}(x_p) - \mathrm{Re}z - i\mathrm{Im}z|} \\ &= \frac{1}{\sqrt{(l_k \mathrm{Re}(x_p) - \mathrm{Re}z)^2 + (l_k \mathrm{Im}(x_p) - \mathrm{Im}z)^2}} \\ &\leq \frac{1}{\delta}.\end{aligned}$$

Finally, since $\delta$ is arbitrary, we can conclude that $f'(z, \Sigma) \asymp g'(z, \widehat{\Sigma})$ for all $z \in D$.

Then let us compute the derivatives. For invertible $A = A(z)$, we have

$$\frac{d(A^{-1})}{dz} = -A^{-1} \frac{dA}{dz} A^{-1},$$

where the derivative is entry-wise. Thus

$$\begin{aligned}&f'(z, T) = -(x_p T - zI)^{-1}(x_p' T - I)(x_p T - z I_p)^{-1} = -(x_p T - z I_p)^{-2}(x_p' T - I), \\ &g'(z, S) = (S - z I_p)^{-2}.\end{aligned}$$

Next, we need to calculate $x' = dx/dz$, where $x(z)$ is the limit of $x_p(z)$. In fact, by looking at the expression of $x_p(z)$, it is not hard to find that $x_p(z)$ is uniformly bounded on $D$. By using a similar argument, we have $x_p' \to x'$ on $D$. To find $x'$, let us start from the following fixed-point equation

$$1 - x = \gamma \left[ 1 + z \mathbb{E}_H \frac{1}{xT - z} \right].$$

Take derivatives on both sides to get

$$\begin{aligned}
-x' &= \gamma \left[ z\mathbb{E}_H \frac{1}{xT - z} \right]' \\
-x' &= \gamma \left[ \mathbb{E}_H \frac{1}{xT - z} + \mathbb{E}_H \frac{z - zTx'}{(xT - z)^2} \right] \\
x' \left[ -1 + \gamma z \mathbb{E}_H \frac{T}{(xT - z)^2} \right] &= \gamma \mathbb{E}_H \frac{xT}{(xT - z)^2} \\
x' &= \frac{\gamma \mathbb{E}_H \frac{xT}{(xT-z)^2}}{-1 + \gamma z \mathbb{E}_H \frac{T}{(xT-z)^2}}.
\end{aligned}$$

Therefore we obtain

$$\begin{aligned}
(\widehat{\Sigma} - zI)^{-2} &\asymp (x_p \Sigma - zI_p)^{-2}(I - x_p' \Sigma) \\
p^{-1} \operatorname{tr}(\widehat{\Sigma} - zI)^{-2} &\asymp -x_p' p^{-1} \operatorname{tr}[\Sigma (x_p \Sigma - zI)^{-2}] + p^{-1} \operatorname{tr}[(x_p \Sigma - zI_p)^{-2}] \\
&\to \frac{\gamma \mathbb{E}_H \frac{xT}{(xT-z)^2}}{1 - \gamma z \mathbb{E}_H \frac{T}{(xT-z)^2}} \mathbb{E}_H \frac{T}{(xT - z)^2} + \mathbb{E}_H \frac{1}{(xT - z)^2} \\
&= \frac{\gamma x \left( \mathbb{E}_H \frac{T}{(xT-z)^2} \right)^2}{1 - \gamma z \mathbb{E}_H \frac{T}{(xT-z)^2}} + \mathbb{E}_H \frac{1}{(xT - z)^2}.
\end{aligned}$$

Now, let $z = -\lambda$ and then we will have

$$\begin{aligned}
(\widehat{\Sigma} + \lambda I)^{-2} &\asymp (x_p \Sigma + \lambda I)^{-2}(I - x_p' \Sigma) \\
p^{-1} \operatorname{tr}(\widehat{\Sigma} + \lambda I)^{-2} &\to \frac{\gamma x \left( \mathbb{E}_H \frac{T}{(xT+\lambda)^2} \right)^2}{1 + \gamma \lambda \mathbb{E}_H \frac{T}{(xT+\lambda)^2}} + \mathbb{E}_H \frac{1}{(xT + \lambda)^2}.
\end{aligned}$$

Finally, we can simply replace $\widehat{\Sigma}, \lambda, \gamma, x$ by $\widehat{\Sigma}_i, \lambda_i, \gamma_i, x_i$ to get the desired results.

3. Limit of $R$: Recall that $R_i = n_i^{-1} \sigma^2 \operatorname{tr}[(\widehat{\Sigma}_i + \lambda_i I_p)^{-2} \widehat{\Sigma}_i]$. We note $p^{-1} \operatorname{tr}(\widehat{\Sigma} + \lambda I)^{-2} \to m_{F_\gamma}'(-\lambda)$ and $\widehat{\Sigma}(\widehat{\Sigma} + \lambda I)^{-2} = (\widehat{\Sigma} + \lambda I)^{-1} - \lambda (\widehat{\Sigma} + \lambda I)^{-2}$, so

$$\frac{\operatorname{tr}[\widehat{\Sigma}(\widehat{\Sigma} + \lambda I)^{-2}]}{n} \to \gamma [m_{F_\gamma}(-\lambda) - \lambda m_{F_\gamma}'(-\lambda)].$$

Hence

$$R_{ii} \to \sigma^2 \left[ \gamma_i [m_{F_{\gamma_i}}(-\lambda_i) - \lambda m'_{F_{\gamma_i}}(-\lambda_i)] \right].$$

Next, we find a limit in terms of population parameters

$$\begin{aligned}
\widehat{\Sigma}(\widehat{\Sigma} + \lambda I)^{-2} &= (\widehat{\Sigma} + \lambda I)^{-1} - \lambda(\widehat{\Sigma} + \lambda I)^{-2} \\
&\asymp (x_p \Sigma + \lambda I)^{-1} - \lambda(x_p \Sigma + \lambda I)^{-2}(I - x'_p \Sigma) \\
p^{-1} \operatorname{tr} \widehat{\Sigma}(\widehat{\Sigma} + \lambda I)^{-2} &\asymp p^{-1} \operatorname{tr}(x_p \Sigma + \lambda I)^{-1} - \lambda p^{-1} \operatorname{tr}\left[(I - x'_p \Sigma)(x_p \Sigma + \lambda I)^{-2}\right] \\
&\to \mathbb{E}_H \frac{1}{xT + \lambda} - \lambda \frac{\gamma x \left( \mathbb{E}_H \frac{T}{(xT+\lambda)^2} \right)^2}{1 + \gamma\lambda \mathbb{E}_H \frac{T}{(xT+\lambda)^2}} - \mathbb{E}_H \frac{\lambda}{(xT + \lambda)^2} \\
&= \mathbb{E}_H \frac{xT}{(xT + \lambda)^2} - \lambda \frac{\gamma x \left( \mathbb{E}_H \frac{T}{(xT+\lambda)^2} \right)^2}{1 + \gamma\lambda \mathbb{E}_H \frac{T}{(xT+\lambda)^2}} \\
&= \frac{x \mathbb{E}_H \frac{T}{(xT+\lambda)^2}}{1 + \lambda\gamma \mathbb{E}_H \frac{T}{(xT+\lambda)^2}},
\end{aligned}$$

where we used the differentiation rule of the calculus of deterministic equivalents. Hence we finally find the limit

$$R_{ii} \to \sigma^2 \left[ \frac{x_i \mathbb{E}_H \frac{T}{(x_i T+\lambda_i)^2}}{1 + \lambda_i \gamma_i \mathbb{E}_H \frac{T}{(x_i T+\lambda_i)^2}} \right].$$

*3.7.8. Proof of Theorem 3.3.2*

Notice that, when the samples are equally distributed and we use the same tuning parameter $\lambda$ for all the local estimators, a direct consequence is that $x_i = x_j = x$ for all $i, j$, where $x$ is the unique solution of the following fixed point equation

$$1 - x = k\gamma \left[ 1 - \lambda \int_0^\infty \frac{dH(t)}{xt + \lambda} \right] = k\gamma \left[ 1 - \mathbb{E}_H \frac{\lambda}{xT + \lambda} \right] = k\gamma(1 - \lambda m_{F_{k\gamma}}(-\lambda)) = k\gamma(1 - \lambda m).$$

In this case, we can express $\mathcal{A}_{ij}$ as

$$\mathcal{A}_{ij} = \sigma^2\alpha^2\mathbb{E}_H\frac{(xT)^2}{(xT+\lambda)^2} = \sigma^2\alpha^2\int\frac{(xt)^2}{(xt+\lambda)^2}dH(t).$$

In order to express $\mathcal{A}_{ij}$ in terms of the sample quantities, we can start from the following equality

$$\int\frac{1}{xt+\lambda}dH(t) = m.$$

Take derivatives with respect to $\lambda$, we have

$$\int\frac{x't+1}{(xt+\lambda)^2}dH(t) = m'.$$

Rearrange terms, we have

$$\int\frac{x't+1}{(xt+\lambda)^2}dH(t) = \int\frac{(xt+\lambda-\lambda)\cdot\frac{x'}{x}+1}{(xt+\lambda)^2}dH(t) = \frac{x'}{x}m+\left(1-\frac{\lambda x'}{x}\right)\int\frac{1}{(xt+\lambda)^2}dH(t) = m'.$$

On the other hand, take derivatives with respect to $\lambda$ on the fixed point equation for $x$ gives us

$$x' = k\gamma(m-\lambda m').$$

So

$$\int\frac{1}{(xt+\lambda)^2}dH(t) = \frac{xm'-x'm}{x-\lambda x'} = \frac{(1-k\gamma)m'+2k\gamma\lambda mm'-k\gamma m^2}{1-k\gamma+k\gamma\lambda^2 m'}.$$

Then we have

$$\begin{aligned}\int\frac{(xt)^2}{(xt+\lambda)^2}dH(t) &= \int\frac{(xt+\lambda-\lambda)^2}{(xt+\lambda)^2}dH(t)\\ &= 1-2\lambda m+\lambda^2\int\frac{1}{(xt+\lambda)^2}dH(t)\\ &= 1-2\lambda m+\lambda^2 m'-\frac{k\gamma\lambda^2(m-\lambda m')^2}{1-k\gamma+k\gamma\lambda m'}.\end{aligned}$$

Now we the expressions for $V, \mathcal{A}, \mathcal{R}$, we also know $\mathcal{W}_k^* = (\mathcal{A}+\mathcal{R})^{-1}V$ and $\mathcal{M}_k = \sigma^2\alpha^2 - V^\top(\mathcal{A}+\mathcal{R})^{-1}V$. By using the auxiliary functions $\mathcal{F}, \mathcal{G}$ defined in the theorem, we have

$$\mathcal{M}_k = \sigma^2\alpha^2 - \frac{1}{\mathcal{G}}V^\top(\mathbf{1}\cdot\mathbf{1}^\top + \mathrm{diag}(\mathcal{F}/\mathcal{G}))^{-1}V = \sigma^2\alpha^2 - \frac{(\sigma^2\alpha^2(1-\lambda m))^2}{\mathcal{G}}\mathbf{1}^\top(\mathbf{1}\cdot\mathbf{1}^\top + \mathrm{diag}(\mathcal{F}/\mathcal{G}))^{-1}\mathbf{1},$$

where $\mathbf{1} = (1, 1, \cdots, 1)^\top$ is the all-one vector. Then similar to the proof of Theorem 3.3.1, we can use the Sherman-Morrison formula to simply the expression, this leads to

$$\begin{aligned}
\mathcal{M}_k &= \sigma^2\alpha^2 - \frac{(\sigma^2\alpha^2(1-\lambda m))^2}{\mathcal{G}}\mathbf{1}^\top\left(\mathrm{diag}(\mathcal{F}/\mathcal{G})^{-1} - \frac{\mathrm{diag}(\mathcal{F}/\mathcal{G})^{-1}\mathbf{1}\cdot\mathbf{1}^\top\,\mathrm{diag}(\mathcal{F}/\mathcal{G})^{-1}}{1+\mathbf{1}^\top\,\mathrm{diag}(\mathcal{F}/\mathcal{G})^{-1}\mathbf{1}}\right)\mathbf{1} \\
&= \sigma^2\alpha^2 - \frac{(\sigma^2\alpha^2(1-\lambda m))^2}{\mathcal{G}}\left(\frac{k\mathcal{G}}{\mathcal{F}} - \frac{k^2\mathcal{G}^2/\mathcal{F}^2}{1+k\mathcal{G}/\mathcal{F}}\right) \\
&= \sigma^2\alpha^2\left(1 - \frac{\sigma^2\alpha^2(1-\lambda m)^2 k}{\mathcal{F}+k\mathcal{G}}\right).
\end{aligned}$$

Similarly, we can express the optimal weights $\mathcal{W}_k^*$ as

$$\begin{aligned}
\mathcal{W}_k^* &= \frac{1}{\mathcal{G}}(\mathbf{1}\cdot\mathbf{1}^\top + \mathrm{diag}(\mathcal{F}/\mathcal{G}))^{-1}V \\
&= \frac{\sigma^2\alpha^2(1-\lambda m)}{\mathcal{G}}(\mathbf{1}\cdot\mathbf{1}^\top + \mathrm{diag}(\mathcal{F}/\mathcal{G}))^{-1}\mathbf{1} \\
&= \frac{\sigma^2\alpha^2(1-\lambda m)}{\mathcal{G}}\left(\mathrm{diag}(\mathcal{F}/\mathcal{G})^{-1} - \frac{\mathcal{G}^2/\mathcal{F}^2}{1+k\mathcal{G}/\mathcal{F}}\mathbf{1}\cdot\mathbf{1}^\top\right)\mathbf{1} \\
&= \frac{\sigma^2\alpha^2(1-\lambda m)}{\mathcal{F}+k\mathcal{G}}\mathbf{1}.
\end{aligned}$$

Why do we need to assume the samples are uniformly distributed across machines? This is a technical assumption. The key difficulty in analyzing MSE and optimal weights comes from the off-diagonal entries of $\mathcal{A}$:

$$\mathcal{A}_{ij} = \sigma^2\alpha^2\mathbb{E}_H\frac{x_i x_j T^2}{(x_i T+\lambda_i)(x_j T+\lambda_j)}.$$

Here, $x_i$ is uniquely determined by the aspect ratio $\gamma_i = p/n_i$, the tuning parameter $\lambda_i$, and

the population distribution $H$. The terms we can handle are of form

$$\mathbb{E}_H \frac{(x_i T)^k}{(x_i T + \lambda_i)^l}, \quad k, l \in \mathbb{N}^*.$$

We can calculate these by taking derivatives of the Stieltjes transform $\mathbb{E}_H \frac{1}{x_i T + \lambda_i}$ and doing further calculations.

We may decompose $\mathcal{A}_{ij}$ into

$$\mathcal{A}_{ij} = \sigma^2 \alpha^2 \left( 1 - \lambda_i \mathbb{E}_H \frac{1}{x_i T + \lambda_i} - \lambda_j \mathbb{E}_H \frac{1}{x_j T + \lambda_j} + \lambda_i \lambda_j \mathbb{E}_H \frac{1}{(x_i T + \lambda_i)(x_j T + \lambda_j)} \right).$$

In general, it is hard to further simplify the last term of the above decomposition. But if we assume the subsample sizes are all equal, then the optimal tuning parameters $\lambda_i$ should also be equal by symmetry. Thus, all $x_i$ are the same and we can rewrite the last term into

$$\lambda_i^2 \mathbb{E}_H \frac{1}{(x_i T + \lambda_i)^2},$$

which is something we can deal with as mentioned above. Besides, from a practical perspective, if the samples are not uniformly distributed, we may need to do much more rounds of communication (compare to Algorithm 1) to find the optimal tuning parameters $\lambda_i$ and the optimal weights $\omega_i$.

*3.7.9. Proof of Theorem 3.4.1*

The proof for $v$ and $R$ is clear by Theorem 3.3.1. For the limit of $A$, the diagonal case is also direct. When $i \neq j$, recall that

$$E_{ij} = p^{-1} \operatorname{tr}\{(\widehat{\Sigma}_i + \lambda_i I_p)^{-1} (\widehat{\Sigma}_j + \lambda_j I_p)^{-1}\} \to \mathbb{E}_H \frac{1}{(x_i T + \lambda_i)(x_j T + \lambda_j)}.$$

For $H = \delta_1$, the expectation decouples, we find

$$E_{ij} \to \frac{1}{x_i + \lambda_i} \cdot \frac{1}{x_i + \lambda_j} = m_{\gamma_i}(-\lambda_i) m_{\gamma_j}(-\lambda_j).$$

Therefore,

$$A_{ij} \to \sigma^2\alpha^2[1-\lambda_i m_{\gamma_i}(-\lambda_i)]\cdot[1-\lambda_j m_{\gamma_j}(-\lambda_j)].$$

Now let us put everything together. Recall that the optimal risk has the form $\mathrm{MSE}^*_{dist} = \|\beta\|^2 - v^\top(A+R)^{-1}v$. Based on the above discussion, we have

$$\sigma^2\alpha^2(A+R) \to \sigma^2\alpha^2(\mathcal{A}+\mathcal{R}) = VV^\top + D,$$

where $D$ is a diagonal matrix with $i$-th diagonal entry $\sigma^2\alpha^2(\mathcal{R}_{ii}+\mathcal{A}_{ii}) - V_i^2$. Then, by using the Sherman−Morrison formula, we have

$$V^\top(VV^\top + D)^{-1}V = \frac{V^\top D^{-1}V}{1+V^\top D^{-1}V}.$$

So the limiting distributed risk is

$$\mathcal{M}_k = \sigma^2\alpha^2 - \sigma^2\alpha^2\frac{V^\top D^{-1}V}{1+V^\top D^{-1}V} = \frac{\sigma^2\alpha^2}{1+V^\top D^{-1}V} = \frac{\sigma^2\alpha^2}{1+\sum_{i=1}^k \frac{V_i^2}{D_i}},$$

which finishes the proof.

*3.7.10. Proof of Proposition 3.4.2*

Recall that

$$\begin{aligned}\frac{V_i^2}{D_i} &= \frac{\sigma^4\alpha^4(1-\lambda_i m_{\gamma_i}(-\lambda_i))^2}{\sigma^4\alpha^4\lambda_i^2[m'_{\gamma_i}(-\lambda_i) - m^2_{\gamma_i}(-\lambda_i)] + \sigma^4\alpha^2\gamma_i[m_{\gamma_i}(-\lambda_i) - \lambda_i m'_{\gamma_i}(-\lambda_i)]} \\ &= \frac{\alpha^2(1-\lambda_i m_{\gamma_i}(-\lambda_i))^2}{\alpha^2\lambda_i^2[m'_{\gamma_i}(-\lambda_i) - m^2_{\gamma_i}(-\lambda_i)] + \gamma_i[m_{\gamma_i}(-\lambda_i) - \lambda_i m'_{\gamma_i}(-\lambda_i)]},\end{aligned}$$

and our goal is to find $\lambda_i$ that maximizes $V_i^2/D_i$. Luckily, from Dobriban and Wager (2018) it follows that for $k=1$, i.e. when there is only one machine, the optimal choice of the tuning parameter $\lambda$ is $\gamma/\alpha^2$. This means that the maximizer of $V^2/D$ is $\lambda = \gamma/\alpha^2$. Now, due to the decoupled structure of $\mathcal{M}_k$, the optimal tuning parameters are $\lambda_i = \gamma_i/\alpha^2$.

Plugging in the parameters, we have

$$\frac{V_i^2}{D_i} = \frac{\alpha^2}{\gamma_i m_{\gamma_i}(-\gamma_i/\alpha^2)} - 1.$$

Then the optimal risk can be simplified to

$$\mathcal{M}_k = \frac{\sigma^2\alpha^2}{1 + \sum_{i=1}^k \left[\frac{\alpha^2}{\gamma_i m_{\gamma_i}(-\gamma_i/\alpha^2)} - 1\right]}.$$

When $k = 1$, this equals to $\sigma^2\gamma m_\gamma(-\gamma/\alpha^2)$ which matches the known result from Dobriban and Wager (2018).

*3.7.11. Proof of Proposition 3.4.3*

The explicit form is easy to derive by plugging $z = -\gamma/\alpha^2$ into the formula of $m_\gamma(z)$. Next, we can check monotonicity by computing $\phi'(\gamma)$:

$$\phi'(\gamma) = \frac{\alpha^2}{2\gamma^2}\left(1 + \frac{(1-1/\alpha^2)\gamma - 1}{\sqrt{[(1-1/\alpha^2)\gamma - 1]^2 + 4\gamma^2/\alpha^2}}\right) > 0.$$

Finally, for the convexity, let us consider the two cases separately.

1. $\alpha \leq 1$: With some effort, we find the second derivative of $\phi$

$$\phi''(\gamma) = \frac{\alpha^2\left(\frac{2\gamma^2}{\alpha^2} - \left(((1-\frac{1}{\alpha^2})\gamma - 1)^2 + \frac{4\gamma^2}{\alpha^2}\right)\left((1-\frac{1}{\alpha^2})\gamma - 1\right)\right) - \left(((1-\frac{1}{\alpha^2})\gamma - 1)^2 + \frac{4\gamma^2}{\alpha^2}\right)^{3/2}\Big)}{\gamma^3[((1-1/\alpha^2)\gamma - 1)^2 + 4\gamma^2/\alpha^2]^{3/2}}.$$

To analyze the second derivative, it is helpful to denote $1 - (1 - \frac{1}{\alpha^2})\gamma$ by $\Delta$. Clearly, in

this case, $\Delta \geq 1$. Then we can rewrite $\phi''$ as

$$\begin{aligned}
\phi''(\gamma) &= \frac{\alpha^2}{\gamma^3[\Delta^2 + 4\gamma^2/\alpha^2]^{3/2}} \left( \frac{2\gamma^2}{\alpha^2} + (\Delta^2 + \frac{4\gamma^2}{\alpha^2})\Delta - (\Delta^2 + \frac{4\gamma^2}{\alpha^2})^{3/2} \right) \\
&= \frac{\alpha^2}{\gamma^3[\Delta^2 + 4\gamma^2/\alpha^2]^{3/2}} \left( \frac{2\gamma^2}{\alpha^2} + (\Delta^2 + \frac{4\gamma^2}{\alpha^2}) \left( \Delta - \sqrt{\Delta^2 + \frac{4\gamma^2}{\alpha^2}} \right) \right) \\
&= \frac{\alpha^2}{\gamma^3[\Delta^2 + 4\gamma^2/\alpha^2]^{3/2}} \left( \frac{2\gamma^2}{\alpha^2} - \frac{4\gamma^2}{\alpha^2} \cdot \frac{\Delta^2 + 4\gamma^2/\alpha^2}{\Delta + \sqrt{\Delta^2 + \frac{4\gamma^2}{\alpha^2}}} \right) \\
&= \frac{\alpha^2}{\gamma^3[\Delta^2 + 4\gamma^2/\alpha^2]^{3/2}} \left( \frac{2\gamma^2}{\alpha^2} - \frac{4\gamma^2}{\alpha^2} \cdot \frac{\sqrt{\Delta^2 + 4\gamma^2/\alpha^2}}{\Delta + \sqrt{\Delta^2 + \frac{4\gamma^2}{\alpha^2}}} \right) \\
&\leq \frac{\alpha^2}{\gamma^3[\Delta^2 + 4\gamma^2/\alpha^2]^{3/2}} \left( \frac{2\gamma^2}{\alpha^2} - \frac{4\gamma^2}{\alpha^2} \cdot \frac{1}{2} \right) = 0.
\end{aligned}$$

Thus, $\phi(\gamma)$ is always concave in this case.

2. $\alpha > 1$: Here we can consider the Taylor expansion of $\phi''$ near the origin. We can check that $\phi''(\gamma) = 2(1 - 1/\alpha^2)\gamma^3 + o(\gamma^3)$ as $\gamma \to 0$, which means $\phi''(\gamma) > 0$ for small $\gamma$. When $\gamma$ is very large, we can immediately see that $\phi''(\gamma) < 0$, since the leading order in the numerator of $\phi''(\gamma)$ is $-\gamma^3$. Then the desired result follows.

*3.7.12. Proof of Theorem 3.4.4*

For the first property, minimizing the ARE is equivalent to maximizing the following quantity

$$\sum_{i=1}^{k} \frac{\gamma_i m_{\gamma_i}(-\gamma_i/\alpha^2)}{\alpha^2} = \sum_{i=1}^{k} \frac{\phi(\gamma_i)}{\alpha^2}.$$

It is helpful to introduce $r(t) = \phi(\gamma)$, where $t = 1/\gamma$. We can easily compute that

$$r'(t) = \frac{\alpha^2}{2} \left( -1 + \frac{t - (1 - 1/\alpha^2)}{\sqrt{(t - (1 - 1/\alpha^2))^2 + 4/\alpha^2}} \right) < 0 \ , r''(t) = \frac{2}{[(t - (1 - 1/\alpha^2))^2 + 4/\alpha^2]^{3/2}} > 0.$$

Thus, $r(t)$ is a decreasing and convex function. We can show the ARE achieves minimum when the samples are equally distributed by considering the following optimization problem

$$\begin{aligned} \max_{t_i} \quad & \sum_{i=1}^{k} \frac{r(t_i)}{\alpha^2} \\ \text{subject to} \quad & \sum_{i=1}^{k} t_i = \frac{1}{\gamma}, \\ & t_i \geq 0, i = 1, 2, \ldots, k. \end{aligned}$$

We denote the objective by $R(t_1, \ldots, t_k)$, and the corresponding Lagrangian by $R_\xi = R - \xi(\sum_i t_i - 1/\gamma)$. Then it is easy to check that the condition $\frac{\partial R_\xi}{\partial t_i} = 0$ reduces to

$$\frac{r'(t_i)}{\alpha^2} - \xi = 0, \ i = 1, 2, \ldots, k.$$

Since $r'(t)$ is also monotone, the unique solution to the stationary condition is $t_1 = t_2 = \cdots = t_k = 1/(k\gamma)$. If some $t_i$ equals to 0, then it reduces to a problem with $k-1$ machines. So it remains to check the boundary case where only one $t_i$ is non-zero and equals to $1/\gamma$. Obviously, this is the trivial case where the ARE is 1. Therefore, we have shown that the ARE attains its minimum when the samples are equally distributed across $k$ machines.

Next, for fixed $\alpha^2$ and $\gamma$, we can check

$$\frac{\partial \psi}{\partial k} = \frac{\gamma m_\gamma(-\gamma/\alpha^2)}{\alpha^2} \left( \frac{\alpha^2}{2\gamma} \cdot \frac{\left(\gamma/\alpha^2 + \gamma\right)^2 k + \gamma/\alpha^2 - \gamma}{\sqrt{\left(\gamma/\alpha^2 + \gamma\right)^2 k^2 + 2\left(\gamma/\alpha^2 - \gamma\right) k + 1}} - \frac{1+\alpha^2}{2} \right) \leq 0.$$

Moreover, the limit of $\psi$ is

$$\begin{aligned} h(\alpha^2, \gamma) = \lim_{k \to \infty} \psi(k, \gamma, \alpha^2) &= \frac{\gamma m_\gamma(-\gamma/\alpha^2)}{\alpha^2} \left( 1 + \frac{\alpha^2}{\gamma(1+\alpha^2)} \right) \\ &= \frac{-\gamma/\alpha^2 + \gamma - 1 + \sqrt{(-\gamma/\alpha^2 + \gamma - 1)^2 + 4\gamma^2/\alpha^2}}{2\gamma} \left( 1 + \frac{\alpha^2}{\gamma(1+\alpha^2)} \right). \end{aligned}$$

Then for fixed $\alpha^2$, we can differentiate $h(\alpha^2, \gamma)$ with respect to $\gamma$:

$$\begin{aligned}\frac{\partial h}{\partial \gamma} = &-\frac{\alpha^2}{\gamma^2(1+\alpha^2)} \cdot \frac{-\gamma/\alpha^2 + \gamma - 1 + \sqrt{(-\gamma/\alpha^2+\gamma-1)^2 + 4\gamma^2/\alpha^2}}{2\gamma} \\ &+ \left(1 + \frac{\alpha^2}{\gamma(1+\alpha^2)}\right) \cdot \frac{1 - 1/\alpha^2 + \frac{(-\gamma/\alpha^2+\gamma-1)(1-1/\alpha^2)+4\gamma/\alpha^2}{\sqrt{(-\gamma/\alpha^2+\gamma-1)^2+4\gamma^2/\alpha^2}}}{2\gamma} \\ &- \left(1 + \frac{\alpha^2}{\gamma(1+\alpha^2)}\right) \cdot \frac{-\gamma/\alpha^2 + \gamma - 1 + \sqrt{(-\gamma/\alpha^2+\gamma-1)^2 + 4\gamma^2/\alpha^2}}{2\gamma^2}.\end{aligned}$$

After tedious calculation, we find $\frac{\partial h}{\partial \gamma} \geq 0$. Finally, we can evaluate the limit of $h$ as $\gamma \to 0$ and $\gamma \to \infty$

$$\lim_{\gamma \to 0} h(\alpha^2, \gamma) = \frac{1}{1+\alpha^2}, \quad \lim_{\gamma \to \infty} h(\alpha^2, \gamma) = 1.$$

On the other hand, for fixed $\gamma$, we can check that $h$ is a decreasing function of $\alpha^2$ and

$$\lim_{\alpha^2 \to 0} h(\alpha^2, \gamma) = 1, \quad \lim_{\alpha^2 \to \infty} h(\alpha^2, \gamma) = \begin{cases} 1 - \frac{1}{\gamma^2}, & \gamma > 1, \\ 0, & 0 < \gamma \leq 1. \end{cases}$$

*3.7.13. Proof of Theorem 3.4.5*

Recall that the optimal weights are $w^* = (A+R)^{-1}v$ and $\sigma^2\alpha^2(A+R) \to VV^\top + D$. Denote the limit of the optimal weights by $W$, so that we have

$$W = \sigma^2\alpha^2(VV^\top + D)^{-1}V = \frac{\sigma^2\alpha^2 D^{-1}V}{1 + V^\top D^{-1}V}.$$

When we choose $\lambda_i = \gamma_i/\alpha^2$ for each $i$, we can write the limiting optimal weights as

$$W = \mathcal{M}_k \cdot D^{-1}V.$$

So, it follows from the formulas of $\mathcal{M}_k, D$ and $V$ that

$$W_i = \left(\frac{\alpha^2}{\gamma_i m_{\gamma_i}(-\gamma_i/\alpha^2)}\right) \cdot \left(\frac{1}{1 + \sum_{i=1}^k \left[\frac{\alpha^2}{\gamma_i m_{\gamma_i}(-\gamma_i/\alpha^2)} - 1\right]}\right).$$

For the sum of the coordinates, we have

$$1^\top W = \frac{\sum_{i=1}^k \left(\frac{\alpha^2}{\gamma_i m_{\gamma_i}(-\gamma_i/\alpha^2)}\right)}{1+\sum_{i=1}^k \left[\frac{\alpha^2}{\gamma_i m_{\gamma_i}(-\gamma_i/\alpha^2)}-1\right]} = \frac{\sum_{i=1}^k \left(\frac{\alpha^2}{\gamma_i m_{\gamma_i}(-\gamma_i/\alpha^2)}\right)}{1-k+\sum_{i=1}^k \left(\frac{\alpha^2}{\gamma_i m_{\gamma_i}(-\gamma_i/\alpha^2)}\right)} \geq 1.$$

In the special case where all $\gamma_i$ are equal, i.e., $\gamma_i = k\gamma$, we have all $W_i$ equal to

$$W_i = \frac{\frac{\alpha^2}{k\gamma \cdot m_{k\gamma}(-k\gamma/\alpha^2)}}{1-k+\frac{\alpha^2}{\gamma \cdot m_{k\gamma}(-k\gamma/\alpha^2)}} = \frac{1}{k+(1-k)\cdot k\gamma/\alpha^2 \cdot m_{k\gamma}(-k\gamma/\alpha^2)}.$$

In terms of the optimal risk function $\phi(\gamma) = \phi(\gamma, \alpha) = \gamma m_\gamma(-\gamma/\alpha^2)$ defined before, this can also be written as the following optimal weight function

$$\mathcal{W}(k, \gamma, \alpha) = \frac{1}{k-(k-1)\cdot \phi(k\gamma)/\alpha^2}.$$

The monotonicity and the limits of $\mathcal{W}$ can be checked directly.

*3.7.14. Proof of Proposition 3.4.6*

Recall the definition of the out-of-sample prediction error is $\mathbb{E}\|y_t - x_t^\top \hat{\beta}\|^2$. So for any estimator $\hat{\beta}$, under the assumption $\Sigma = I$, we have

$$\begin{aligned}
\mathbb{E}\|y_t - x_t^\top \hat{\beta}\|^2 &= \mathbb{E}\|x_t^\top(\hat{\beta}-\beta)+\varepsilon_t\|^2 = \mathbb{E}\|x_t^\top(\hat{\beta}-\beta)\|^2 + \sigma^2 \\
&= \mathbb{E}[(\hat{\beta}-\beta)^\top x_t \cdot x_t^\top(\hat{\beta}-\beta)] + \sigma^2 \\
&= \mathbb{E}[(\hat{\beta}-\beta)^\top \Sigma(\hat{\beta}-\beta)] + \sigma^2 \\
&= \mathbb{E}\|\hat{\beta}-\beta\|^2 + \sigma^2.
\end{aligned}$$

When we consider the distributed estimator and take the limit, we obtain

$$\mathcal{O}_k = \sigma^2 + \mathcal{M}_k,$$

and the formula for OE. For the inequality between OE and ARE, it is sufficient to notice that $ARE \leq 1$. Finally, the explicit formulas follow easily from previous results.

*3.7.15. Proof of Theorem 3.4.7*

It is equivalent to show that the ARE is always greater than or equal to $1/(1+\alpha^2)$. To do this, we need to use Theorem 3.4.4. From the first property, we have ARE$\geq \psi(k, \gamma, \alpha^2)$. Then, since $\psi$ is a decreasing function of $k$, it is lower bounded by its limit at infinity, which is $h(\alpha^2, \gamma)$. Finally, $h(\alpha^2, \gamma)$ is an increasing function of $\gamma$, so it is lower bounded by the limit at 0, which is $1/(1+\alpha^2)$. When $\gamma > 1$, $h(\alpha^2, \gamma)$ is a decreasing function of $\alpha^2$, so it is lower bounded by the limit at infinity, which is $1 - 1/\gamma^2$. The desired result follows.

# CHAPTER 4 : Selecting the Number of Components in PCA via Random Signflips

This chapter is based on Hong et al. (2020), which is a joint work with Dr. David Hong and my advisor Professor Edgar Dobriban. I contributed to a large portion of ideas, derivations, and simulations.

## 4.1. Introduction

Discovering latent low-dimensional phenomena in large and messy datasets is one of the central challenges faced in modern data analysis. Indeed, examples arise across virtually all of science and engineering, and unsupervised dimensionality reduction is a standard component in statistical analysis. In particular, Factor Analysis (FA) and Principal Component Analysis (PCA) remain incredibly popular and successful techniques. They continue to be integral parts of myriad data analysis pipelines, being performed routinely in thousands of studies every year. Applications abound in psychology and education (Horn, 1965; Tran and Formann, 2009), public health (Patil et al., 2010), management/marketing (Stewart, 1981), economics/finance (Bai and Ng, 2002; Ahn and Horenstein, 2013), genomics (Lin et al., 2016; Yano et al., 2019), environmental sensing (Subbarao et al., 1996), and manufacturing (Apley and Shi, 2001), to name just a few. See, e.g., Anderson (2003); Jolliffe (2002); Yao et al. (2015), for references.

Given measurements of $p$ features (covariates) over a set of $n$ samples (datapoints), FA and PCA identify common factors driving variation in the data. However, these components do not all capture *meaningful* variation, i.e., signal; many capture variation simply due to noise. Hence, an important question is: how many components capture signals rising above the noise? This paper tackles this challenge in the increasingly common (but as yet relatively unaddressed) setting where the noise can be heterogeneous. Methods that do not appropriately account for heterogeneity can dramatically degrade, and theory developed for homogeneous cases do not directly apply. New methods and theory are needed.

*4.1.1. Selecting the number of factors from noisy data*

This paper centers on the important problem of selecting the number of factors from data with heterogeneous noise. Informally, we are given data that is modelled as being a sum of signal and noise

$$X = S + N \in \mathbb{R}^{n \times p}, \quad r := \operatorname{rank} S \ll n.$$

We wish to estimate how many of the leading principal components of $X$ capture the signal $S$ rather than the noise $N$. The noise entries $N$ are random with potentially heterogeneous distributions. The rank $r$ of $S$ can provide a reasonable upper bound, but some components of $S$ may be too small, and can get "buried" in the noise. Moreover, since $r$ is unknown, proper statistical methods are needed to estimate the number of components.

Estimating how many factors to keep is well known to significantly impact downstream data analyses, with the standard textbook Brown (2014) calling it "the most crucial decision" in exploratory FA. Choosing too few deprives downstream steps of potentially critical information, while choosing too many passes on unnecessary noise. Moreover, data in many important applications have weak "emergent" factors that are nontrivial to identify, making this a challenging problem. Such settings are common, e.g., in behavioral and biological sciences. Consequently, much work has gone into the development of many methods. Indeed, there are many more than can be discussed in detail here so we instead give a brief high-level overview.

Classical and standard methods to factor selection include the scree plot (Cattell, 1966; Cattell and Vogelmann, 1977), i.e., Cattell's scree plot, sphericity tests based on likelihood ratios (Bartlett, 1954; Lawley, 1956), the minimum average partial test (Velicer, 1976), and approaches based on minimum description length (Wax and Kailath, 1985; Fishler et al., 2002). A popular and practical choice among classical methods is parallel analysis (Horn, 1965; Buja and Eyuboglu, 1992). Owen and Wang (2016) note that "there is a large amount of evidence that PA is one of the most accurate [...] classical methods for determining the number of factors". Indeed many works find PA to be highly effective; see,

e.g., the discussion in Dobriban (2020, Section 1.2) and references therein.

More recently, tremendous progress has been made by using modern insights from high-dimensional probability and random matrix theory to study large-dimensional data. For settings with strong factors, methods based on information criteria were studied by Bai and Ng (2002); Alessi et al. (2010); Bai et al. (2018). Kapetanios (2004, 2010) considered pure white noise from a random matrix theory perspective. Onatski (2010) argued for using the differences between adjacent eigenvalues, and Lam and Yao (2012); Wang (2012); Ahn and Horenstein (2013) analogously proposed using the ratio. Spike models with diverging spikes were also considered in Cai et al. (2020), and Kaiser (1960); Fan et al. (2020) studied the correlation matrix. For settings with weak "emergent" factors, which are our primary focus, Nadakuditi and Edelman (2008) study an information criterion-based method, Kritchman and Nadler (2009) uses a hypothesis test connected with Roy's largest root test. Passemier and Yao (2014); Ke et al. (2020) study spiked models under various assumptions, and Owen and Wang (2016) propose a bi-cross-validation approach. See Fan et al. (2014); Johnstone and Paul (2018) and references therein for more details. Indeed, great strides have been made on developing and analyzing rigorous methods, fueled by modern theoretical insights.

However, much work to date has been for homogeneous noise, and such techniques can dramatically degrade when noise is heterogeneous (as we show below for parallel analysis). The analysis of large-dimensional data with heterogeneous noise is an actively developing area. In this paper, we use modern insights from random matrix theory to develop and analyze an elegant variant of the popular and practical classical parallel analysis method that carefully accounts for heterogeneous noise.

*4.1.2. Our contributions*

This paper proposes a new variant of the popular and practical parallel analysis (PA) method for the increasingly important modern setting of data with heterogeneous noise. We consider a general "signal-plus-noise" model for large-dimensional data, where the noise matrix has independent entries with heterogeneous variances. This is sometimes called a

model with a "general variance profile", and has received recent attention in random matrix theory, see e.g., (Girko, 2001; Hachem et al., 2006, 2008; Husson, 2020). The standard spiked covariance model for PCA and the popular linear factor model are both special cases. However, statistically rigorous methods for selecting the number of components have not yet been proposed for the general model. We make the following main contributions:

1. **New method: Signflip Parallel Analysis.** We propose the new *Signflip Parallel Analysis (Signflip PA)* method for selecting the number of components/factors (i.e., the rank) in the general "signal-plus-noise" model. It is a type of parallel analysis method. It compares the singular values of the data (or equivalently, the eigenvalues of the sample covariance matrix) to those of "empirical null" data generated by randomly, independently and uniformly flipping the signs of the data matrix entries. The selected rank is the number of leading data singular values that rise above their signflipped analogues, where the comparison is done sequentially starting from the top singular value and stopping at the first failure.

2. **Theoretical characterization in signal-plus-noise models.** By extending the framework developed in Dobriban (2020), we characterize Signflip PA by analyzing the ability of signflips to (a) "destroy" the signals, i.e., the operator norm of signflipped signals vanish, while (b) "preserving" the noise. This allows us to conclude that Signflip PA consistently selects the number of above-noise factors. We need to extend the framework of Dobriban (2020), because signflips do not in general preserve the distribution exactly; for this we extend the framework to only require "consistent noise level estimation". This is a much more broadly applicable condition, and it requires the powerful tools described below. See theorem 4.5.3 for our main result in factor models.

   (a) **Signal destruction.** We develop elegant sufficient conditions for asymptotic signal destruction that reveal the importance of signal delocalization and rank. The first set of conditions applies to general signal matrices, while the second set exploits the special structure of sums of outer products (which commonly arise in practice).

Moreover, we derive necessary conditions that match the sufficient conditions for signals with uniformly bounded rank, i.e., we find *necessary and sufficient* conditions for this important case.

We discovered that we can derive these conditions by building on and extending recent random matrix theory breakthroughs on dimension-free operator norm bounds for heterogeneous random matrices (Latała et al., 2018). To our knowledge, these specific results (specifically the ones depending on the so-called logarithmic decay coefficient discussed below) have not yet been used in any application in statistics, machine learning, or data science. Thus, this approach can be viewed as a theoretical innovation of our work.

(b) **Noise level estimation.** We prove that Signflip PA asymptotically consistently estimates the correct overall "noise level". This is equivalent to saying that it recovers the leading singular values of the underlying noise under a heterogeneous noise model with general variance profile. We extend the framework of Dobriban (2020), showing that recovery in this sense is sufficient. Full invariance of the joint noise distribution is not needed, allowing us to handle a broad class of noise distributions. The proof also leverages very recent results on large deviations for the top singular value of random matrices with variance profiles (Husson, 2020), which, to our knowledge, have not been used either in statistical applications until now.

3. **Theoretical justification for signflips.** Signflip PA naturally suggests considering a broader class of "wild bootstrap"-like methods that destroy signals by multiplying each entry with an independent random variable. However, we show that random signflips have a certain special justification in this setting.

4. **Theoretical explanation for the degradation of Permutation PA.** We explain why Permutation PA is not effective for heterogeneous noise. Roughly speaking, permutations

homogenize the noise and fail to recover the underlying noise level, which can lead to severe over- or under-estimation of the rank. We make this intuition precise by showing that the spectrum of permuted noise converges to a generalized Marchenko-Pastur law parameterized by column-wise averaged (i.e., homogenized) variances. Notably, the random matrix of interest (permuted noise) has dependent entries for which we extend a technique developed for correlated random matrices (Bai and Zhou, 2008).

5. **Implications for rank selection.** Finally, we explain the implications of the above general signal and noise results for rank selection, in the special cases of factor analysis and PCA. We show that Signflip PA is asymptotically consistent for selecting the number of perceptible factors in certain general linear factor models, which includes the popular spiked models for PCA. Our theoretical conditions allow both growing numbers and strengths of factors. Even in the special case where Permutation PA is applicable, they are strictly more general than the previous results from Dobriban (2020).

6. **Empirical support.** We empirically validate the theory and method through a broad range of numerical simulations and experiments. We find Signflip PA to be accurate in a wide set of simulated data models, matching Permutation PA for homogeneous noise while remaining effective for heterogeneous noise. Moreover, Signflip PA performs well, including compared to standard methods, on both realistically generated chlorine data and empirical single cell RNA-sequencing data.

The structure of our paper follows the above outline, with most proofs in the appendix.

### 4.2. Parallel analysis, heterogeneous noise, and the need for new methods

Here we provide needed background on parallel analysis and heterogeneous noise. We conclude with the observation that Permutation PA is incredibly effective and successful for homogeneous noise, but can be dramatically inaccurate for heterogeneous noise. New methods are needed.

*4.2.1. Parallel analysis via permutations*

Parallel analysis (PA), introduced by Horn (1965), and its permutation version (Buja and Eyuboglu, 1992) are among the most popular methods for rank selection. The key idea is that we expect components rising above the noise to produce data singular values rising above their "null" pure-noise analogues. Generating "parallel" datasets, e.g., via column-wise permutation, gives estimates of these null singular values, providing data-driven cut-offs.

The permutation version, which we call *Permutation PA*, is described in algorithm 3 and we illustrate it for a rank-one example in fig. 18. First we form the $n \times p$ data matrix

---

**Algorithm 3:** Parallel analysis via permutations (Permutation PA)
(Buja and Eyuboglu, 1992)

---

**Input :** Data matrix $X \in \mathbb{R}^{n \times p}$ ($n$ samples and $p$ features), percentile $\alpha$, number of trials $T$.

**Output:** Selected number of factors $\hat{k}$.

$\sigma \leftarrow$ singular values of $X$;

**for** $t \leftarrow 1$ **to** $T$ **do**

  Randomly permute entries in each column of $X$:

$$X_\pi = \begin{pmatrix} X_{\pi_1^{(1)} 1} & \cdots & X_{\pi_p^{(1)} p} \\ \vdots & & \vdots \\ X_{\pi_1^{(n)} 1} & \cdots & X_{\pi_p^{(n)} p} \end{pmatrix},$$

  where $\pi_1, \ldots, \pi_p$ are permutations of $(1, \ldots, n)$ drawn independently and uniformly at random;

  $\tilde{\sigma}^{(t)} \leftarrow$ singular values of $X_\pi$;

**end**

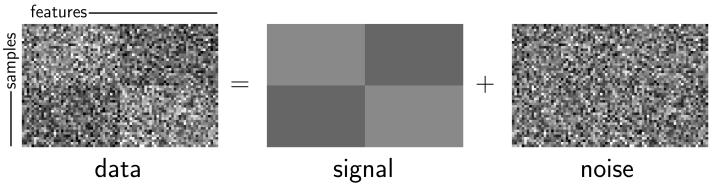
$\hat{k} \leftarrow$ first $k$ for which either of the two conditions below holds

$$\sigma_{k+1} \leq \alpha\text{-percentile of } \left\{ \tilde{\sigma}_{k+1}^{(1)}, \ldots, \tilde{\sigma}_{k+1}^{(T)} \right\}, \qquad \text{(pairwise)}$$

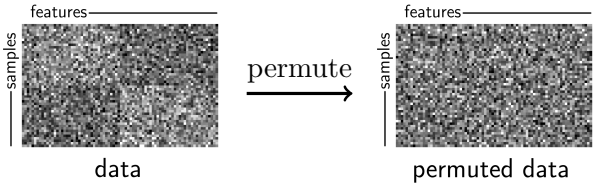
$$\sigma_{k+1} \leq \alpha\text{-percentile of } \left\{ \tilde{\sigma}_{1}^{(1)}, \ldots, \tilde{\sigma}_{1}^{(T)} \right\}, \qquad \text{(upper-edge)}$$

i.e., $\hat{k}$ is the number of leading singular values above the $\alpha$-percentile of their permuted analogues, where "pairwise" and "upper-edge" are two choices for the comparison.

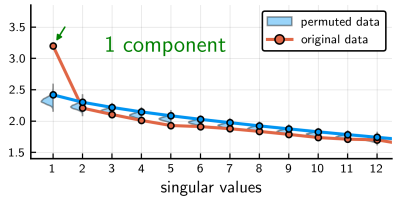
---




(a) Rank-one signal in homogeneous noise.




(b) Shuffle data by permuting each column indep.




(c) Select "above-permuted" factors.

Figure 18: Illustration of Permutation PA for a rank-one signal in homogeneous noise. Permutations scramble signal structure, creating a noise-like matrix. PA selects those factors whose singular values rise above a chosen percentile of their permuted analogues, in this case correctly selecting one factor.

$X$ ($n$ samples and $p$ features). In our example this is a rank-one signal in homogeneous noise as shown in fig. 18a. Then we shuffle each column of $X$ independently by randomly permuting its entries, forming permuted data $X_\pi$ as shown in fig. 18b. Each column has a random permutation independent from all other columns. Repeating this $T$ times and collecting the singular values from each trial, we form an empirical (marginal) distribution for each singular value of $X_\pi$, as shown in fig. 18c. The procedure concludes by selecting the number of leading data singular values that rise above a chosen percentile $\alpha$ (e.g., 50%, 95% or 100%) of their permuted analogues. In our example this correctly selects one component. This is a pairwise and sequential criterion, shown in algorithm 3 as eq. (pairwise). We start from the first data singular value, including it if it is larger than the $\alpha$ percentile of the first permuted-data singular value, and moving on to the second singular values, and so on. We stop the first time this criterion fails.

While this *pairwise comparison* is the classical method, another popular alternative in practice is to compare all data singular values against the $\alpha$ percentile of only the first permuted singular value. Indeed, some recent methods even aim to directly estimate the noise upper-edge (Dobriban and Owen, 2018). We will distinguish this method by

calling it *upper-edge comparison* since it compares against the largest (or upper-edge) of the permuted singular values. Algorithm 3 shows this option as eq. (upper-edge). This method has the benefit of only requiring the calculation and storage of the first singular value, i.e., the operator norm $\|X_\pi\|$, of the permuted data. Moreover, it provides a more conservative threshold and less frequently over-selects, as we see in the numerical experiments of sections 4.6.1 and 4.6.2.

In fig. 18b, observe that the rank-one block structure visible in the original data becomes lost in the permuted data. This instead looks noise-like, and Permutation PA correctly selects one factor. Though this example is intentionally simple to focus on illustrating the procedure, the same occurs in general. Permutation PA is incredibly effective in practice, with numerous endorsements and increasing popularity among applied statisticians, especially in the biological sciences (e.g., Brown, 2014; Lin et al., 2016). Furthermore, it is a natural method that can be intuitively understood without appealing to sophisticated theoretical tools. Taken together with its simplicity (only a few lines of code!), Permutation PA is an incredibly attractive method in practice as well as a great foundational tool to study and build upon.

Buja and Eyuboglu (1992) provided some basic theoretical justification from the perspective of hypothesis testing, working under a certain null distribution. Assuming independent and identically distributed (i.i.d.) samples, i.e. rows, the permutation distribution of the matrix is the conditional null distribution under a non-parametric null $H_0$ of complete independence, conditioning on the minimal sufficient statistic under $H_0$. The minimal sufficient statistic is the $p$-tuple of empirical distributions of each column. Viewed through this lens, Permutation PA is a sort of quasi-inferential method. A theoretical justification under large-dimensional signal-plus-noise models was developed in recent years by Dobriban (2020). Using tools from random matrix theory, this work rigorously analyzed how permutations destroy signal structure while preserving the noise, providing a precise explanation as well as sufficient conditions for consistent selection by Permutation PA of so-called "above-noise"

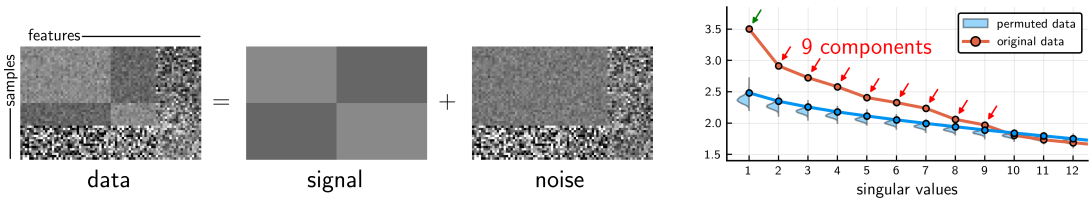
factors. Building on these new insights, Dobriban and Owen (2018) proposed a deterministic variant. Continuing theoretical insights into the application of parallel analysis ideas for modern large-dimensional settings is an exciting and burgeoning research front; see, e.g., Zhou (2019); McKennan (2020); Chen and Li (2020); Fan et al. (2020). In this paper, we take a step towards further developing and using these insights to improve robustness to the heterogeneity we expect to become an increasingly common part of modern data analysis.

*4.2.2. Data with heterogeneous noise and the need for new techniques*

Heterogeneous noise arises very naturally in modern settings, whether due to heteroskedasticity in the features or due to heterogeneous quality among samples. For example, the noise level in medical imaging varies both within images and from image to image. Likewise, atmospheric corruptions in astronomical data vary both from night to night and from pixel to pixel, and the quality of environmental sensors can vary from location to location. These types of effects all contribute to heterogeneity in the noise. Moreover, we expect such heterogeneity to only become more common, especially as datasets are increasingly built up from samples collected at myriad and varying places, times or by varying equipment.

Consequently, recent works have begun to study how to properly account for heterogeneous noise when carrying out PCA for large data. Much work centers on improving the quality of the estimated components by, e.g., correcting for bias due to heterogeneity *across features* (Zhang et al., 2018), or by using an optimal spectral shrinkage after whitening the noise (Leeb and Romanov, 2018). Other methods include optimal denoising with respect to losses that account for heterogeneity (Leeb, 2019), or optimal weighting of samples to account for heterogeneity *across samples* (Hong et al., 2016, 2018a,b). While much work remains, indeed great progress has already been made. However, fewer works have addressed the question of *how to estimate* the number of components in these heterogeneous settings.

For heterogeneity across features, Leeb and Romanov (2018) consider selecting the singular values rising (at least slightly) above the asymptotic operator norm of the noise matrix. This can be predicted when the noise is whitened or when the noise variances are well-




(a) Rank-one signal in heterogeneous noise. (b) Selection by Permutation PA.

Figure 19: Illustration of Permutation PA applied to a rank-one signal in *heterogeneous* noise. PA underestimates the noise level and dramatically over-selects as a consequence.

estimated. Ke et al. (2020) consider a setting where these noise variances are drawn from a Gamma distribution, and propose exploiting this knowledge by first fitting the Gamma distribution from the bulk singular values. The problem of rank selection without exact or distributional knowledge of the variances, when noise is heterogeneous across both features and samples, has remained relatively open and is the setting of our work.

One might hope to use existing approaches, such as Permutation PA, that are rigorously grounded and battle-tested in the homogeneous setting. However, it turns out that Permutation PA can severely underperform in these heterogeneous settings. We illustrate this in fig. 19 for a simple rank-one signal in heterogeneous noise. In this example, the noise variance varies throughout the data. The data matrix is moderately noisy on the right side, least noisy in the upper left, and most noise in the lower left. Permutation PA is not effective here. It incorrectly selects nine factors, over-selecting by eight. Moreover, this loss in performance is not unique to fig. 19, and can happen in general when the noise is heterogeneous. Roughly speaking, permutations "smear" and homogenize the noise in a way we will make precise in section 4.4.5, where we also provide a detailed explanation and characterization of this phenomenon.

Summarizing, while Permutation PA is incredibly effective and enjoys many theoretical guarantees in the homogeneous setting, it is far less so in the increasingly important case where noise is heterogeneous. New methods and theory are needed.


(a) Signflip each data entry indep. with prob. 1/2. (b) Select "above-signflip" factors.

Figure 20: Illustration of Signflip PA for the heterogeneous example of fig. 19, for which Permutation PA incorrectly selected nine components. Signflip PA accurately recovers the noise and correctly selects one component.

### 4.3. Proposed method: Signflip parallel analysis

The dramatic degradation of Permutation PA under heterogeneous noise could naturally lead one to consider abandoning the approach in this setting. However, we propose an elegant and simple modification that largely retains the excellent performance of parallel

---

**Algorithm 4:** Parallel analysis via signflips (Signflip PA)

---

**Input :** Data matrix $X \in \mathbb{R}^{n \times p}$ ($n$ samples and $p$ features), percentile $\alpha$, number of trials $T$.

**Output:** Selected number of factors $\hat{k}$.

$\sigma \leftarrow$ singular values of $X$;

**for** $t \leftarrow 1$ **to** $T$ **do**

  Randomly signflip entries of $X$: form $R \circ X$ where

$$R_{ij} \overset{iid}{\sim} \begin{cases} +1, & \text{with probability } 1/2, \\ -1, & \text{with probability } 1/2, \end{cases}$$

  i.e., $R \in \mathbb{R}^{n \times p}$ has independent identically distributed Rademacher entries;

  $\tilde{\sigma}^{(t)} \leftarrow$ singular values of $R \circ X$;

**end**

$\hat{k} \leftarrow$ first $k$ for which either

$$\sigma_{k+1} \leq \alpha\text{-percentile of } \left\{ \tilde{\sigma}_{k+1}^{(1)}, \ldots, \tilde{\sigma}_{k+1}^{(T)} \right\}, \qquad \text{(pairwise)}$$

$$\sigma_{k+1} \leq \alpha\text{-percentile of } \left\{ \tilde{\sigma}_{1}^{(1)}, \ldots, \tilde{\sigma}_{1}^{(T)} \right\}, \qquad \text{(upper-edge)}$$

i.e., $\hat{k}$ is the number of leading singular values above the $\alpha$-percentile of their signflipped analogues, where "pairwise" and "upper-edge" are two choices for the comparison.

---

analysis under homogeneous noise, while expanding these benefits to data with heterogeneous noise.

Specifically, we propose replacing random permutations with random *entrywise signflips*. We call the resulting method *Signflip PA*. For clarity, we describe the full procedure in algorithm 4, but note that it is essentially the same as Permutation PA (algorithm 3) except for line 3. Now we generate signflipped data $R \circ X$, where $\circ$ denotes the Hadamard (entrywise) product, and $R \in \{\pm 1\}^{n \times p}$ has i.i.d. Rademacher entries, i.e., each entry $R_{ij}$ is $+1$ or $-1$ with equal probability. Put another way, we flip the sign of each entry with probability one-half, independent of the rest.

Figure 20 illustrates the process for the same heterogeneous data as fig. 19, for which Permutation PA incorrectly selected nine components. As shown in fig. 20a, we begin by independently signflipping each entry of the data matrix $X$, forming signflipped data $R \circ X$. Observe how the rank-one signal that can be visually seen in the data is no longer visible after signflipping, while the heterogeneous noise profile remains. The remainder proceeds analogously to Permutation PA. We repeat $T$ times and collect the singular values from each trial, forming an empirical (marginal) distribution for each singular value of $R \circ X$, as shown in fig. 20b. Finally, we select the number of leading data singular values that rise above the $\alpha$ percentile of their signflipped analogues in the pairwise and sequential way stated in algorithm 4 as eq. (pairwise). Signflip PA correctly selects one component.

As before, we also consider the popular method of *upper-edge comparison* eq. (upper-edge) that compares all data singular values against (the $\alpha$ percentile of) only the first signflipped singular value. Which selection rule to choose tends to depend on the application and the salient priorities. Recall that upper-edge comparison never selects more factors than pairwise comparison, making it more conservative (see examples in sections 4.6.1 and 4.6.2). Moreover, upper-edge comparison has the benefit of only requiring us to calculate and store the first singular value, i.e., the operator norm $\|R \circ X\|$, of the signflipped data. The two selection rules turn out to be essentially asymptotically equivalent as $n, p \to \infty$ for fixed $k$,

and in many settings they agree. Indeed, with either selection rule, Signflip PA correctly selects one component in this example.

The simplicity of just replacing permutations with signflips is one of the standout features of Signflip PA, as it immediately inherits many of the same practical benefits enjoyed by Permutation PA. It is an equally natural method that can be easily digested and understood without appealing to sophisticated theoretical results. It is also easy to implement, again taking only a few lines of code. These features highlight the benefit of building on the framework of parallel analysis. A thorough characterization of its performance does indeed require nontrivial theoretical work; our analysis in section 4.4 leverages recent breakthroughs in random matrix theory. However, even the overall ability of signflips to preserve heterogeneous noise may be believable at an intuitive level given that each noise entry is treated separately.

Signflip PA also admits a natural interpretation as a sort of quasi-inferential method when viewed through the lens of independence testing. Assume independent but not identically distributed rows. Let $H_0$ be the null that the columns are independent and the marginal entry distributions are all symmetric about zero. Then, the minimal sufficient statistic becomes the $n \times p$ array of absolute values $|X_{ij}|$ and the signflip distribution is the conditional null distribution. See also Bordenave et al. (2020) for different uses of sign-flips in sparse matrix completion. Section 4.4 analyzes Signflip PA under signal-plus-noise models by extending the framework developed in Dobriban (2020). We show that Signflip PA is able to "destroy" low-rank signals in very general settings by estimating the noise level. The main implication for factor models is given in Theorem 4.5.3.

## 4.4. Theoretical analysis and guarantees

This section gives theoretical insight to answer the important question: how does Signflip PA work, and when does it work in general? Building on the framework developed in Dobriban (2020), we analyze general signal-plus-noise models and characterize when signflipping: a) "destroys" low-rank signal structure, and b) "recovers" heterogeneous noise.

We will make all these notions precise below, but the overall intuitive picture is shown in

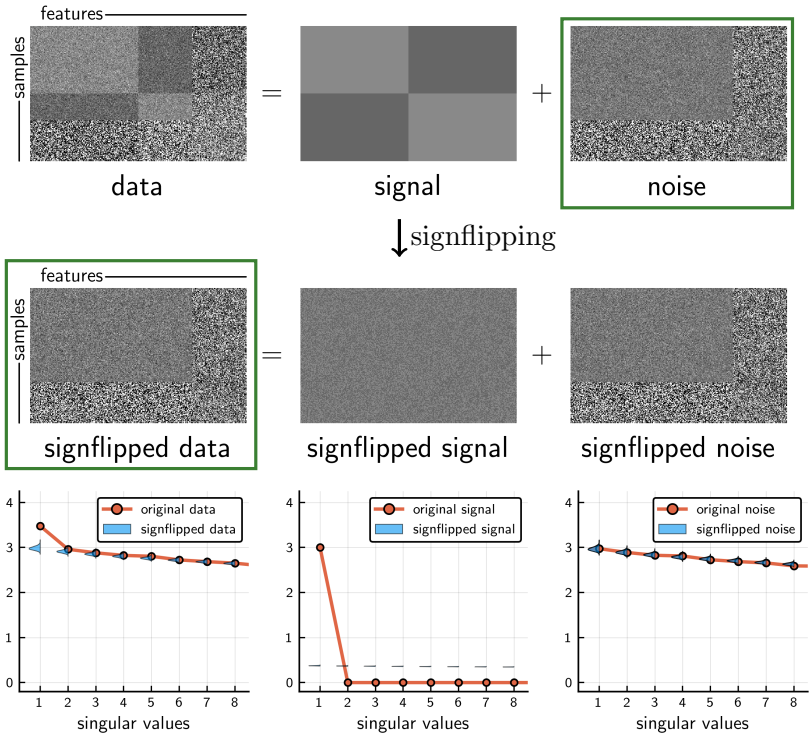



Figure 21: Preview and rough intuition for theoretical analysis. Signflipping "destroys" low-rank signals (in operator norm) and consistently estimates the noise level—the singular values of signflipped data $R \circ X$ are close to those of the noise $N$.

fig. 21. The underlying low-rank signal structure is scrambled by signflipping (producing a matrix with much smaller operator norm), while the signflipped noise is essentially indistinguishable from the original noise (and has very similar singular values). As a result, the signflipped data looks like the noise (including its heterogeneous variance profile), and in particular has very similar singular values. Our theoretical analysis makes these rough observations rigorous.

After some background and notational clarifications (section 4.4.1), we begin by characterizing signal destruction by signflips (section 4.4.2), followed by an analysis of corresponding noise level estimation (section 4.4.3). Finally, we explain why signflips are uniquely suited (section 4.4.4) and in what way permutations homogenize heterogeneous noise (section 4.4.5). Carefully leveraging recent breakthroughs in random matrix theory enable us to obtain elegant and simple conditions throughout.

*4.4.1. Notations and preliminaries*

To make the following discussions precise, we detail our notations and provide some relevant theoretical background here.

**Notations.** Throughout the paper, we denote the Hadamard product (entrywise multiplication) of two matrices $A$ and $B$ of the same size by $A \circ B$. For a $m \times n$ matrix $A$, we use $\|A\|$ and $\|A\|_F$ to denote the spectral norm and the Frobenius norm, respectively. Let $|A|$ be the matrix whose $(i,j)$-th entry is the absolute value of the $(i,j)$-th entry of $A$. Let $\|A\|_p$ denote the matrix norms induced by vector norms, and $\|A\|_{p,q}$ denote the entrywise matrix norms. They are defined as follows

$$\|A\|_p = \sup_{x \neq 0} \frac{\|Ax\|_p}{\|x\|_p}, \quad \|A\|_{p,q} = \left( \sum_{j=1}^{n} \left( \sum_{i=1}^{m} |a_{ij}|^p \right)^{q/p} \right)^{1/q},$$

where $\|x\|_p$ denote the $p$-norm for vectors. The $(2,\infty)$ norm $\|A\|_{2,\infty}$ will play a special role; it is the maximum of the $\ell_2$ norms of the columns of $A$. Similarly, $\|A^\top\|_{2,\infty}$ is the max of the $\ell_2$ norms of the rows of $A$. We also denote the Schatten $k$-norm of a matrix $A$ by $\|A\|_{S_k}$. Letting $\sigma_i(A)$ be the $i$-th largest singular value of $A$, the Schatten $k$-norm is defined as

$$\|A\|_{S_k} = \left( \sum_i |\sigma_i(A)|^k \right)^{1/k}.$$

We denote the trace of a matrix $A$ by $\mathrm{tr}(A)$. For two random matrices $A, B$, $A =_d B$ means that the matrices have the same distribution, thus implying that the corresponding $(i,j)$-entries of $A$ and $B$ have the same distribution. We use the classical big-O and little-o notations to describe the asymptotic relationship between two quantities. We call a random variable $Y$ a Rademacher random variable if $\mathbb{P}(Y = -1) = \mathbb{P}(Y = 1) = 1/2$. We use $f \lesssim g$ to denote that $f \leq Cg$ for a universal constant $C$ which does not depend on any parameter of the problem unless stated explicitly. We will use $f \asymp g$ if $f \lesssim g$ and $g \lesssim f$.

**Statistical model.** The model we will consider in this paper is the following. The $n \times p$ data matrix $X$ has $n$ samples and $p$ features. The rows of $X$ are independent $p$-dimensional observations, not necessarily identically distributed. We can express $X$ in the following "signal-plus-noise" form:

$$X = S + N.$$

Here $S$ is the "signal" part, which is typically of low rank. We denote the unknown $\text{rank}(S) = r \ll \min(n, p)$; this is the key parameter we aim to estimate. The "noise" part $N$ is modelled as $N = n^{-1/2}(T \circ E)$, where $E$ has i.i.d. random entries with zero mean, unit variance, and finite fourth moment, $T$ is a deterministic matrix with $(i, j)$-entry $T_{ij}$. Thus, $N$ has independent entries and the $(i, j)$-entry has variance $T_{ij}^2/n$. We say that $N$ has a *general variance profile*, where the profile matrix is $T$. This model is a generalization of the standard factor model. In the standard factor model, within each column of $N$, all entries have the same variance.

Define the aspect ratio of $X$ as $\gamma_p = p/n$. We will work in the proportional limit regime (e.g., Marchenko and Pastur, 1967; Serdobolskii, 2007; Johnstone, 2007; Yao et al., 2015, etc), where we consider a sequence of problems with growing parameters $n, p \to \infty$ such that $\gamma_p \to \gamma \in (0, \infty)$ as $n, p \to \infty$. For a positive semidefinite matrix $A$, let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ be its eigenvalues, and its empirical spectral distribution be defined as

$$F_A(x) = \frac{1}{p} \sum_{i=1}^{p} \mathbf{1}(\lambda_i \leq x).$$

As usual, we will typically assume $F_A(x)$ converges weakly to a limiting spectral distribution $H$. For a non-square matrix, we can still define its empirical spectral distribution, by using its singular values instead. For a bounded probability distribution $H$, we define its upper edge to be

$$\mathcal{U}(H) := \inf\{M \in \mathbb{R} | H(M) = 1\}.$$

**Random matrix theory.** Now, we will briefly talk about some needed results from random

matrix theory (RMT). See (Bai and Silverstein, 2010; Couillet et al., 2011; Yao et al., 2015) for references. We assume the $n \times p$ design matrix $X$ is generated as $X = Z\Psi^{1/2}$ for an $n \times p$ matrix $Z$ with i.i.d. entries, satisfying $\mathbb{E}(Z_{ij}) = 0$, $\mathbb{E}(Z_{ij}^2) = 1$, and $\mathbb{E}(Z_{ij}^{4+\varepsilon}) < \infty$. The empirical spectral distribution of the $p \times p$ positive semidefinite matrix $\Psi$ has a limiting spectral distribution, in the sense of weak convergence. Under these assumptions, a central result in this area is the Marchenko-Pastur theorem (Marchenko and Pastur, 1967; Bai and Silverstein, 2010), which says that the empirical spectral distribution of the sample covariance matrix $n^{-1}X^\top X$ converges weakly to a limiting spectral distribution $F = F_{\gamma,H}$ almost surely as $p/n \to \gamma, n, p \to \infty$. Moreover, the largest eigenvalue of $n^{-1}X^\top X$ will also converge almost surely to the upper edge of $F$. A common approach to prove this type of result is to use the *Stieltjes transform*. For a probability distribution $F$ over $\mathbb{R}$, the Stieltjes transform $m_F(z)$ of $F$ is a complex analytic function defined as

$$m_F(z) = \int_{-\infty}^{\infty} \frac{1}{x - z} dF(x), \quad \forall z \in \mathbb{C} \setminus \operatorname{supp}(F).$$

An important property of the Stieltjes transform that $m_F$ uniquely determines $F$. The intuition behind this approach is that for a symmetric matrix $A \in \mathbb{R}^{p \times p}$, the Stieltjes transform of its empirical spectral distribution $F_A$ is

$$m_{F_A}(z) = \int \frac{1}{x - z} dF_A(x) = \frac{1}{p} \sum_{i=1}^{p} \frac{1}{\lambda_i(A) - z} = \frac{1}{p} \operatorname{tr} (A - zI)^{-1}.$$

Thus, in order to study the convergence of the empirical spectral distribution $F_A$, we can work with the Stieltjes transform $m_{F_A}(z)$ instead, which boils down to work with the resolvent matrix $(A - zI)^{-1}$. For $\operatorname{tr}(A - zI)^{-1}$, there are many matrix inversion lemmas and matrix identities we can use. Similar results based on singular values also hold for non-square matrices.

*4.4.2. Signal destruction by signflips*

Here we describe our results on signal destruction by Signflip PA, needed for the general theory of consistent signal selection. One might wonder in what sense signflipping "destroys"

the signal, given that there is no reduction in Frobenius norm, i.e., $\|R \circ S\|_F = \|S\|_F$. In other words, the sum-of-squares of the singular values are unchanged. The key is that signflipping takes low-rank signals (for which this sum is dominated by the first few singular values) and makes them "noise-like" (with the energy spread out among all singular values). Consequently, the signal is *destroyed in operator norm*: $\|R \circ S\| \to 0$.

This section proves sufficient as well as necessary conditions for signals $S$ guaranteeing that $\|R \circ S\| \to 0$ as $n, p \to \infty$, either in $L^1$ or almost surely. Recall that $L^1$ convergence and almost sure convergence both imply convergence in probability. We provide conditions for general signal matrices as well as sums of outer products (which are common in many applications). We finally show that our conditions are in fact optimal for signals with uniformly bounded rank. The conditions we find for Signflip PA are generally simpler and sharper than those found for Permutation PA (Dobriban, 2020), even for homogeneous noise. This is because we are able to build on recent breakthroughs and a deep understanding of heterogeneous random matrices with independent entries.

**General conditions guaranteeing signal destruction**

We begin with our most general conditions for signal destruction, which build on an extensive line of works on the operator norm of random matrices with independent heterogeneous Gaussian entries (e.g., Latała, 2005; Bandeira and Van Handel, 2016; Latała et al., 2018, and references therein). In particular, the major breakthrough Latała et al. (2018) characterizes the precise dimension-free behavior of the Schatten norms of these matrices. We adapt it to our setting by relating this to the operator norm of signflipped matrices, and build on it by deriving bounds in terms of the signal rank.

We need the following decay coefficient (which we referred to as the "logarithmic decay coefficient" above), measuring the rate of decay of the row and column norms:

$$\rho_\infty(X) := \max_{i=1,\dots,m+q} \left\| \begin{matrix} X \\ X^\top \end{matrix} \right\|_{\infty,(i)} \sqrt{\log i}. \tag{4.1}$$

Here $\|A\|_{\infty,(i)}$ denotes the $i$-th largest column $\ell_\infty$ norm, i.e., $\|A\|_{\infty,(1)} \geq \cdots \geq \|A\|_{\infty,(q)}$ sorts the column norms $\|A_{:1}\|_\infty, \ldots, \|A_{:q}\|_\infty$ in descending order. Intuitively, if the row and column norms of $X$ decay quickly, then $\rho_\infty(X)$ is small.

We will assume that the rows and columns of $S$ have asymptotically vanishing $\ell_2$ norms in expectation, which turns out to be necessary (section 4.4.2). One can verify that if they do not vanish, then the operator norm of $R \circ S$ cannot converge to zero (consider the canonical basis vectors to get a lower bound). We allow both random and deterministic signals.

**Theorem 4.4.1** (Asymptotic signal destruction)**.** *Let $S = S_{n,p} \in \mathbb{R}^{n\times p}$ be a sequence of signal matrices, and let $R = R_{n,p} \in \{\pm 1\}^{n\times p}$ be a sequence of Rademacher random matrices of corresponding size. Suppose that $S$ has asymptotically vanishing column/row norms in expectation: $\mathbb{E}\|S\|_{2,\infty} \to 0$ and $\mathbb{E}\|S^\top\|_{2,\infty} \to 0$. Then we have as $n, p \to \infty$,*

*$L^1$ **convergence:** $\mathbb{E}\|R \circ S\| \to 0$ if additionally either:*

*(a) the magnitude signal $|S|$ decays in expected operator norm: $\mathbb{E}\||S|\| \to 0$,*

*(b) the decay coefficient eq. (4.1) vanishes in expectation: $\mathbb{E}\{\rho_\infty(S)\} \to 0$, or*

*(c) the expected largest column/row norms vanish fast enough:*
*$\mathbb{E}\|S\|_{2,\infty} = o\{\log^{-1/4}(n+p)\}$ and $\mathbb{E}\|S^\top\|_{2,\infty} = o\{\log^{-1/4}(n+p)\}$.*

*Moreover, sufficient condition (b) is guaranteed under any of the following conditions:*

- *the $\ell_k$ norm of the entries of $S$ vanishes: $\mathbb{E}\|S\|_{k,k} \to 0$ for some $k \geq 2$ (not necessarily an integer),*
- *$\mathbb{E}\{\mathrm{rank}^{1/2}(S)\sqrt{\|S\|_{2,\infty} \cdot \|S^\top\|_{2,\infty}}\} \to 0$, or*
- *$\mathrm{rank}(S)$ is uniformly bounded.*

***Almost sure convergence:*** *$\|R \circ S\| \to_{a.s.} 0$ if there exists $k \geq 2$ (not necessarily an integer) for which $\mathbb{E}\{\|S\|_{2,k}^k + \|S^\top\|_{2,k}^k\}$ is summable (over $n, p$).*

*When $S$ is deterministic, the expectations with respect to $S$ are dropped.*

This theorem (proved in section 4.7.2) provides general conditions under which signal destruction is guaranteed by random signflips. Recall that all these conditions are also sufficient conditions for convergence *in probability*. Roughly speaking, we require either a small signal (i.e., vanishing magnitude operator norm) or sufficient delocalization across rows and columns.

*Remark* 1. Signals with uniformly bounded rank (which is a common assumption) automatically have sufficient delocalization under the assumption of vanishing row/column $\ell_2$ norms. This provides a necessary and sufficient condition for such signals. We formalize and elaborate on this fact in section 4.4.2.

*Remark* 2. Sufficient condition (a) for $L^1$ convergence may appear simple, leading one to wonder if it is implied by either of the other two. However, this is *not* the case. Consider $S = \log^{\alpha}(2n) \cdot I_{n \times n}$ with $\alpha \in (-\frac{1}{4}, 0)$. One can verify $\|S\|_{2,\infty} = \|S^\top\|_{2,\infty} = \log^{\alpha}(2n) \to 0$ and $\||S|\| = \log^{\alpha}(2n) \to 0$, and indeed $\mathbb{E}\|R \circ S\| = \log^{\alpha}(2n) \to 0$ (in fact, this is deterministically true). However,

$$\rho_\infty(S) = \log^{\alpha+1/2}(2n) \to \infty,$$
$$\log^{1/4}(2n)\|S\|_{2,\infty} = \log^{1/4}(2n)\|S^\top\|_{2,\infty} = \log^{\alpha+1/4}(2n) \to \infty.$$

Hence we see that sufficient condition (a) is not redundant. It captures signals that do not delocalize per se and essentially vanish on their own.

*Remark* 3. For clarity and convenience, we state most of our results in the large matrix limit as $n, p \to \infty$, as this setting is our primary focus. However, one can verify that many of our results, especially in theorem 4.4.1, do not strictly require this and generalize immediately to arbitrary sequences of signal matrices (e.g., with only $n$ growing).

### Conditions for sums of outer products

While theorem 4.4.1 is quite powerful and general, it is also very useful to consider signals $S$ written as sums of outer products, i.e.,

$$S = \theta_1 u_1 v_1^\top + \cdots + \theta_r u_r v_r^\top \in \mathbb{R}^{n \times p},$$

as these arise very naturally in practice. Some important examples are:

- The singular value decomposition (SVD) $S = \sum_{i=1}^r \theta_i u_i v_i^\top$, where $\theta_1, \ldots, \theta_r$ are singular values with corresponding orthonormal sets of left and right singular vectors $u_i$ and $v_i$.
- $S$ is random where $u_i$ and/or $v_i$ are independent random vectors. This is the setting for standard factor models Anderson (2003); Brown (2014) and certain spiked PCA models, e.g., Benaych-Georges and Nadakuditi (2012); Couillet and Debbah (2011); Yao et al. (2015); Johnstone and Paul (2018), etc.

We do not require these terms to be orthogonal, nor even linearly independent. We will also later allow the number of terms $r$ (which upper bounds the rank of $S$) to grow with $n, p$, where we typically consider the setting where $p/n \to \gamma \in (0, \infty)$. To simplify the presentation, however, we start by characterizing a single outer product.

**Theorem 4.4.2** (Signal destruction for an outer product)**.** *Let $S = S_{n,p} = \theta u v^\top \in \mathbb{R}^{n \times p}$ be a sequence of outer product signals with deterministic signal strength $\theta$ and independent signal vectors $u$ and $v$ normalized so that $\mathbb{E}\|u\|_2 = \mathbb{E}\|v\|_2 = 1$, and let $R = R_{n,p} \in \{\pm 1\}^{n \times p}$ be a sequence of Rademacher random matrices of corresponding size. Then we have as $n, p \to \infty$*

*$L^1$ **convergence:** $\mathbb{E}\|R \circ S\| \to 0$ if $\theta \cdot \mathbb{E}(\|u\|_\infty + \|v\|_\infty) \to 0$.*

***Almost sure convergence:** $\|R \circ S\| \to_{a.s.} 0$ if there exists $k \geq 2$ (not necessarily an integer) for which $\theta^k \cdot \mathbb{E}(\|u\|_k^k \cdot \|v\|_2^k + \|u\|_2^k \cdot \|v\|_k^k)$ is summable (over $n, p$).*

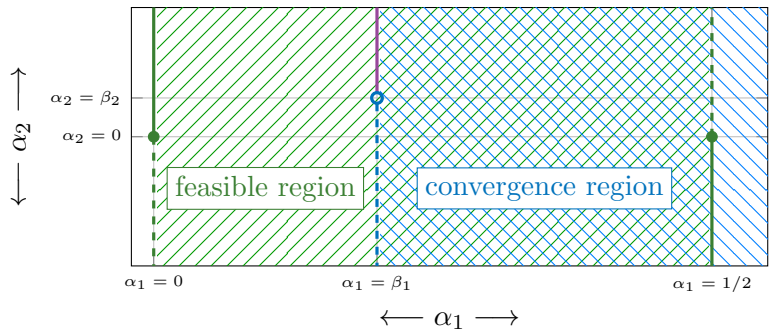



Figure 22: Regimes for delocalization rates $\mathbb{E}\|u\|_\infty, \mathbb{E}\|v\|_\infty = O(p^{-\alpha_1} \log^{-\alpha_2} p)$ for given signal strength rates $\theta = O(p^{\beta_1} \log^{\beta_2} p)$ from theorem 4.4.3: feasible range (north-east green), $L^1$ convergence (north-west blue and purple), and almost sure convergence (north-west blue).

*If the signal vectors $u$ and $v$ are also deterministic, the above expectations are dropped.*

The theorem is proved in section 4.7.3.

*Remark* 4. The normalization $\mathbb{E}\|u\|_2 = \mathbb{E}\|v\|_2 = 1$ is not strictly necessary, but it turns out to be convenient for simplifying some of the expressions. It also provides a natural signal representation, making it possible to reason by rough analogy to the SVD. Removing or modifying the normalization produces fairly similar statements.

The condition for signal destruction simplifies dramatically in this case (note that the rank is uniformly bounded), and it depends only on how fast $\mathbb{E}\|u\|_\infty$ and $\mathbb{E}\|v\|_\infty$ decay compared to the growth of the signal strength $\theta$. The following corollary quantifies these rates, revealing an elegant characterization for both $L^1$ and almost sure convergence.

**Corollary 4.4.3** (Conditions in terms of signal strength and delocalization rates)**.** *Under the setting of theorem 4.4.2, suppose that the signal $S = \theta uv^\top \in \mathbb{R}^{n \times p}$ grows at a rate of $\theta = O(p^{\beta_1} \log^{\beta_2} p)$ and delocalizes at rates of $\mathbb{E}\|u\|_\infty = O(p^{-\alpha_1} \log^{-\alpha_2} p)$ and $\mathbb{E}\|v\|_\infty = O(p^{-\alpha_1} \log^{-\alpha_2} p)$. Then as $n, p \to \infty$ with $p/n \to \gamma$, we have*

*$L^1$ **convergence:** $\mathbb{E}\|R \circ S\| \to 0$ if either: a) $\alpha_1 > \beta_1$, or b) $\alpha_1 = \beta_1$ and $\alpha_2 > \beta_2$.*

***Almost sure convergence:** $\|R \circ S\| \to_{a.s.} 0$ if $S$ is deterministic and $\alpha_1 > \beta_1$.*

The corollary is proved in section 4.7.4.

*Remark* 5. This parameterization is convenient because it covers many important settings. For example, when the singular vectors $u$ and $v$ are independent random vectors uniformly distributed on the unit sphere, it follows that $\mathbb{E}\|u\|_\infty, \mathbb{E}\|v\|_\infty = O(p^{-1/2}\log^{1/2} p)$. One can verify this fact from, e.g., Vershynin (2018, Exercise 2.5.10 and Theorem 3.4.6).

Figure 22 illustrates the convergence regions as a function of the *delocalization exponents* $\alpha_1$ and $\alpha_2$ given *signal growth exponents* $\beta_1$ and $\beta_2$. These exponents are constrained as shown by the feasible region in fig. 22 due to the following simple bounds:

$$1 = \mathbb{E}\|u\|_2 \geq \mathbb{E}\|u\|_\infty \geq \frac{1}{\sqrt{n}}\mathbb{E}\|u\|_2 = \frac{1}{\sqrt{n}} \sim \frac{\sqrt{\gamma}}{\sqrt{p}},$$
$$1 = \mathbb{E}\|v\|_2 \geq \mathbb{E}\|v\|_\infty \geq \frac{1}{\sqrt{p}}\mathbb{E}\|v\|_2 = \frac{1}{\sqrt{p}}.$$

Namely, the feasible range is $\alpha_1 \in (0, 1/2)$ unless $\alpha_2 \geq 0$ for which $\alpha_1 = 0$ is feasible, or $\alpha_2 \leq 0$ for which $\alpha_1 = 1/2$ is feasible.

If the signal strength decays, i.e., $\beta_1 < 0$, all feasible delocalization exponents result in signal destruction (both in $L^1$ and almost surely) as one might expect. This can be quickly verified by observing that the convergence region completely covers the feasible region in fig. 22. On the other hand, if the signal grows too rapidly, i.e., $\beta_1 > 1/2$, there is no overlap and none of the feasible delocalization exponents satisfy our conditions for signal destruction. Indeed, it turns out that the signal is not destroyed in this case (see section 4.4.2 for discussion of necessary conditions). For $u$ and $v$ generated independently uniformly on the unit sphere, signal destruction in $L^1$ occurs as long as $\beta_1 < 1/2$ or $\beta_1 = 1/2$ with $\beta_2 < -1/2$, and occurs almost surely as long as $\beta_1 < 1/2$.

We now generalize theorem 4.4.2 to general sums of outer products, where the number of terms $r$ may even grow in $n, p$. The proof is given in section 4.7.5.

**Theorem 4.4.4** (Signal destruction for a sum of outer products)**.** *Let* $S = S_{n,p} = \sum_{i=1}^{r} \theta_i u_i v_i^\top \in$

$\mathbb{R}^{n\times p}$ *be a sequence of signals, each a sum of* $r = r_{n,p}$ *outer products with deterministic signal strengths* $\theta_1, \dots, \theta_r$ *and left vectors* $u_1, \dots, u_r$ *independent from right vectors* $v_1, \dots, v_r$*, all normalized so that* $\mathbb{E}\|u_i\|_2 = \mathbb{E}\|v_i\|_2 = 1$*. Let* $R = R_{n,p} \in \{\pm 1\}^{n\times p}$ *be a sequence of Rademacher random matrices of corresponding size. Then we have*

$L^1$ ***convergence:*** $\mathbb{E}\|R \circ S\| \to 0$ *if* $\sum_{i=1}^r \theta_i \cdot \mathbb{E}(\|u_i\|_\infty + \|v_i\|_\infty) \to 0$.

***Almost sure convergence:*** $\|R \circ S\| \to_{a.s.} 0$ *if there exists* $k \geq 2$ *(not necessarily an integer) for which* $\mathbb{E}\big\{ \sum_{i=1}^r \theta_i (\|u_i\|_k^k \cdot \|v_i\|_2^k + \|u_i\|_2^k \cdot \|v_i\|_k^k)^{1/k} \big\}^k$ *is summable.*

*If the signal vectors* $u_i$ *and* $v_i$ *are also deterministic, the above expectations are dropped.*

As before signal destruction roughly occurs when the signal vectors delocalize at a rate outpacing the overall growth of the signal strength. As before, we quantify these rates, where we now additionally suppose the number of terms grows as $r = O(p^{\nu_1} \log^{\nu_2} p)$, where we call $\nu_1$ and $\nu_2$ the *rank growth exponents*. Note that technically the rank of $S$ may be lower than $r$ due to the potential for linear dependence among the terms. The following corollary shows how rank can grow in this more general setting; the proof is given in section 4.7.6.

**Corollary 4.4.5** (Conditions in terms of signal rank, strength, and delocalization rates)**.** *Under the setting of theorem 4.4.4, suppose the signal* $S = \sum_{i=1}^r \theta_i u_i v_i^\top \in \mathbb{R}^{n\times p}$ *has rank growing as* $r = O(p^{\nu_1} \log^{\nu_2} p)$ *and signal strength growing as* $\max_i \theta_i = O(p^{\beta_1} \log^{\beta_2} p)$*. Also suppose the signal* $\ell_\infty$ *norms are bounded as*

$$\max_i \mathbb{E}\|u_i\|_\infty = O(p^{-\alpha_1} \log^{-\alpha_2} p), \qquad \max_i \mathbb{E}\|v_i\|_\infty = O(p^{-\alpha_1} \log^{-\alpha_2} p).$$

*Then as* $n, p \to \infty$ *with* $p/n \to \gamma$*, we have*

$L^1$ ***convergence:*** $\mathbb{E}\|R \circ S\| \to 0$ *if we have: a)* $\alpha_1 > \nu_1 + \beta_1$*, or b)* $\alpha_1 = \nu_1 + \beta_1$ *and* $\alpha_2 > \nu_2 + \beta_2$.

***Almost sure convergence:*** $\|R \circ S\| \to_{a.s.} 0$ *if* $S$ *is deterministic and* $\alpha_1 > \nu_1 + \beta_1$.

The rank effectively inflates the signal strength since signflips must now destroy all terms in the sum, which requires a greater amount of delocalization. This produces a trade-off between the signal growth and the rank growth; they cannot both grow rapidly. In many applications $\nu_1 = \nu_2 = 0$, i.e., the rank of the signal is uniformly bounded. This is a common setting in factor analysis and PCA. In this case, the conditions essentially reduce to the rank-one case. However, our theory allows for much more general settings.

**Necessary conditions**

This section establishes some necessary conditions for Signflip PA, i.e., properties the signal $S$ must have to be destroyed by random signflips.

**Theorem 4.4.6** (Necessary conditions for asymptotic signal destruction)**.** *Let $S = S_{n,p} \in \mathbb{R}^{n\times p}$ be a sequence of signal matrices, and let $R = R_{n,p} \in \{\pm 1\}^{n\times p}$ be corresponding Rademacher random matrices. We can have $\mathbb{E}\|R \circ S\| \to 0$ only if the column/row norms vanish in $L^1$: $\mathbb{E}\|S\|_{2,\infty} \to 0$ and $\mathbb{E}\|S^\top\|_{2,\infty} \to 0$. Likewise, only if the following expected matrix norms vanish:*

$$\mathbb{E}\|S\|_{\infty,\infty}, \frac{1}{\sqrt{n}}\mathbb{E}\|S\|_F, \frac{1}{\sqrt{p}}\mathbb{E}\|S\|_F, \frac{1}{\sqrt{n}}\mathbb{E}\|S\|_1, \frac{1}{\sqrt{p}}\mathbb{E}\|S\|_\infty \to 0.$$

For the reader's benefit we provide a complete proof in section 4.7.7. The result follows largely from standard properties of matrix norms with the following general observation.

**Lemma 4.4.7** (Sign-invariant operator norm bounds)**.** *Let $f(X) \geq 0$ be a sign-invariant lower bound on the operator norm (up to a constant), i.e., $f(X) = f(|X|)$ and $f(X) \lesssim \|X\|$. If $\mathbb{E}\|R \circ X\| \to 0$ for Rademacher random matrices $R$, then $\mathbb{E}f(X) \to 0$.*

The lemma gives a general recipe for deriving necessary conditions. It follows immediately from the following observation:

$$f(X) = f(|X|) = f(|R \circ X|) = f(R \circ X) \lesssim \|R \circ X\|.$$

As with the sufficient conditions before, we also provide necessary conditions for sums of outer products, specifically for deterministic signals in SVD form. The proof is in section 4.7.8. One might hope that favorable cancellation among terms might help with signal destruction. We find that this is not the case for the SVD. Essentially, each term must undergo signal destruction.

**Corollary 4.4.8** (Necessary conditions for destruction of an SVD)**.** *Let* $S = S_{n,p} = \sum_{i=1}^{r} \theta_i u_i v_i^\top \in \mathbb{R}^{n\times p}$ *be a sequence of deterministic signals in SVD form with rank* $r = r_{n,p}$*, singular values* $\theta_1, \dots, \theta_r$*, left vectors* $u_1, \dots, u_r$*, and right vectors* $v_1, \dots, v_r$*. Let* $R = R_{n,p} \in \{\pm 1\}^{n\times p}$ *be corresponding Rademacher random matrices. Then* $\mathbb{E}\|R \circ S\| \to 0$ *only if*

$$\max\left[\max_i \theta_i \|u_i\|_\infty,\ \max_i \theta_i \|v_i\|_\infty,\ \frac{1}{\min(n,p)} \sum_{i=1}^{r} \theta_i^2\right] \to 0.$$

**An optimal condition for bounded rank signals**

Determining whether the sufficient conditions are also necessary is a hard question in general. However, for the important setting of signals with uniformly bounded rank (common for low-rank models), we discover a remarkably simple condition for signal destruction (in $L^1$) that is both necessary and sufficient. It is a direct consequence of theorem 4.4.1 and theorem 4.4.6.

**Theorem 4.4.9** (Necessary and sufficient condition for signals with uniformly bounded rank)**.** *Let* $S = S_{n,p} \in \mathbb{R}^{n\times p}$ *be a sequence of signals with uniformly bounded rank, i.e.,* $\operatorname{rank} S = O(1)$*, and let* $R = R_{n,p} \in \{\pm 1\}^{n\times p}$ *be the corresponding Rademacher random matrices. Then* $\mathbb{E}\|R \circ S\| \to 0$ *if and only if the column/row norms vanish in* $L^1$*:* $\mathbb{E}\|S\|_{2,\infty} \to 0$ *and* $\mathbb{E}\|S^\top\|_{2,\infty} \to 0$*.*

In particular, we find a complete characterization for the expected operator norm of signflipped bounded rank signals, namely:

$$\mathbb{E}\|R \circ S\| \asymp \|S\|_{2,\infty} + \|S^\top\|_{2,\infty},$$

which may be of independent interest. Characterizing the expected operator norm of heterogeneous Rademacher random matrices beyond bounded rank heterogeneity remains an open problem.

*4.4.3. Noise level estimation by signflips*

Having analyzed when signflips destroy low-rank signals in operator norm, we now turn to the estimation of the noise level by signflips. We briefly discuss the case covered by Permutation PA, which Dobriban (2020) studied by considering the strong condition of *noise invariance*. Namely, $N_\pi =_d N$, where the equality in distribution is taken with respect to both the noise $N$ and independent column-wise permutations $\pi$. In that case, one can allow noise of the form

$$N = n^{-1/2}(\mathcal{E}D^{1/2} + \mathbf{1}z^\top \Sigma^{1/2}).$$

Here $\mathcal{E}$ is a $n \times p$ matrix of i.i.d. standard Gaussians, $D$ is diagonal, $z \sim \mathcal{N}(0, I_p)$, and $\Sigma$ is a $p \times p$ PSD matrix. The term $\mathbf{1}z^\top \Sigma^{1/2}$ adds a per-column-fixed random variable to each entry. This is allowed by the theory, but it is rarely of practical interest. Thus, we will consider noise models of the form $N = n^{-1/2}\mathcal{E}D^{1/2}$. We also need the convergence of the operator norm: $\|N\| \to b > 0$ as $n, p \to \infty$, which is guaranteed by Proposition 4.2 of Dobriban (2020). Essentially, Permutation PA works well when the noise is homogenous in the sense that different rows (samples) have the same variance within each column (feature). This is the standard model used in factor analysis.

Signflip PA also works for this model. Gaussian random variables $Z$ are symmetric, i.e., $Z =_d -Z$, so it follows that $R \circ N =_d N$ for any fixed signflip matrix $R$, and likewise for random $R$ (independent of $\mathcal{E}$). However, Signflip PA also works beyond this noise model. Suppose the noise matrix $N$ has independent normal entries with heterogeneous variances. Then we say $N$ has a *general variance profile*. Clearly, we still have $R \circ N =_d N$ so signflips continue to be effective.

How about relaxing the Gaussianity assumption on the noise entries? For Permutation PA,

this is not a problem because $N_\pi =_d N$ still holds even when the entries of noise are not Gaussian random variables. But for Signflip PA, when the noise entries are not symmetric random variables, we do not have $R \circ N =_d N$ in general. This may appear to be an issue for Signflip PA at the first glance. However, due to the well-known *universality* phenomenon in random matrix theory (e.g., Tao and Vu, 2011; Erdős et al., 2011), it turns out that Signflip PA can also work beyond Gaussian entries. The key idea is that *sign invariance* can be replaced with a weaker notion of *noise level estimation* that is sufficient for our purposes. We explain this below. We begin with the following definition.

**Definition 4.4.10.** *We say that a random variable $Z$ has a sharp sub-Gaussian Laplace transform (Guionnet and Husson, 2020) if*

$$\mathbb{E} \exp(tZ) \leq \exp\left(\frac{t^2 \mathbb{E}(Z^2)}{2}\right), \forall t \in \mathbb{R}.$$

*We will sometimes say (as shorthand) that $Z$ is a sharp sub-Gaussian random variable.*

*Remark* 6. The term "sharp" comes from the observation that if a random variable $Z$ is sub-Gaussian for some constant $b > 0$,

$$\mathbb{E} \exp(tZ) \leq \exp\left(\frac{b^2 t^2}{2}\right), \forall t \in \mathbb{R},$$

then $\mathbb{E}(Z) = 0$ and $\mathbb{E}(Z^2)$=Var$(Z) \leq b^2$. Some simple examples are of sharp sub-Gaussian random variables are centered Gaussian random variables, Rademacher random variables, and uniform random variables on $[-\sqrt{3}, \sqrt{3}]$. We can generate more complex examples using that for any $t \in [0, 1]$, $\sqrt{t}X + \sqrt{1-t}Y$ has a sharp sub-Gaussian Laplace transform when $X, Y$ are independent sharp sub-Gaussian random variables. One can refer to Guionnet and Husson (2020) for more details.

Then, we have the following theorem, based on results from Girko (2001); Couillet and Debbah (2011); Husson (2020). The recent results in Husson (2020), as far as we know, have not yet been used in any application to statistics. We apply it to show that for noise

$N$ with a general variance profile, there is a limiting spectral distribution. Moreover, the largest singular value of $N$ converges to the supremum of the support of the limiting spectral distribution. Thus, Signflip PA can preserve the limiting spectral distribution as well as the limit of the largest singular value. This result is essential, since it provides a rigorous justification for Signflip PA under heterogenous noise models.

**Theorem 4.4.11** (Sign-invariant heterogenous noise models)**.** *Let $N = n^{-1/2}(T \circ E) \in \mathbb{R}^{n\times p}$, where $E = E_{n,p} \in \mathbb{R}^{n\times p}$ has independent sharp sub-Gaussian entries with zero mean and unit variance, and $T = T_{n,p} \in \mathbb{R}^{n\times p}$ is one of the following variance profiles ($N_{ij}$ has variance $T_{ij}^2/n$):*

***Piecewise constant variance profile:*** *Let $\{\alpha_0, \dots, \alpha_k\}$ and $\{\beta_0, \dots, \beta_l\}$ be two partitions of $[0,1]$ such that $0 = \alpha_0 < \cdots < \alpha_k = 1$ and $0 = \beta_0 < \cdots < \beta_l = 1$, where $k, l$ are fixed. Denote $\tau : [0,1]^2 \to \mathbb{R}_{\geq 0}$ the piecewise constant function defined by $\tau(x,y) = \tau_{st}$ if $x \in [\alpha_{s-1}, \alpha_s)$ and $y \in [\beta_{t-1}, \beta_t)$, where $s = 1, \dots, k$ and $t = 1, \dots, l$. Then, consider $T$ defined by $T_{ij} = \tau(i/n, j/p)$.*

***Continuous variance profile:*** *Let $\tau : [0,1]^2 \to \mathbb{R}_{\geq 0}$ be a continuous function. Suppose $T$ satisfies*

$$\lim_{n,p\to\infty} \sup_{i,j} |T_{ij} - \tau(i/n, j/p)| = 0.$$

*Then the noise $N$ is sign-invariant, i.e., $|\sigma_k(R \circ N) - \sigma_k(N)| \to_{a.s.} 0$ as $n, p \to \infty$ with $p/n \to \gamma$ for any $k$ (fixed w.r.t. $n, p$) where $R = R_{n,p} \in \{\pm 1\}^{n\times p}$ is the corresponding Rademacher random matrix.*

*In particular, the empirical spectral distributions of $N$ and $R \circ N$ both converge weakly to a deterministic distribution $\bar{F}$ with probability one, and $\sigma_k(N), \sigma_k(R \circ N) \to_{a.s.} \tilde{\sigma}$ for any fixed $k$, where $\tilde{\sigma}$ is the rightmost point in the support of $\bar{F}$.*

The proof is given in section 4.8.1. In conclusion, the largest singular value of the true noise $N$ and the signflipped noise $R \circ N$ have the same limiting value. This shows that Signflip

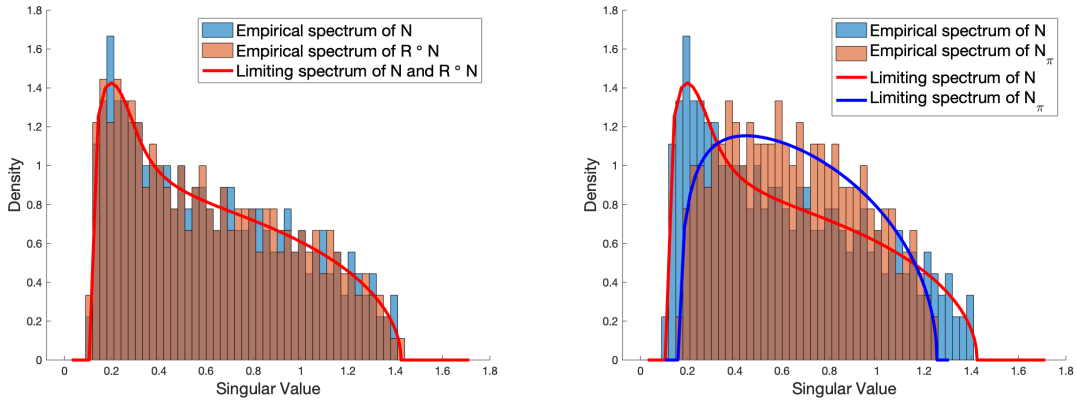
PA asymptotically estimates the proper noise level, which is the limit of the top true noise singular value. This implies that it uses the correct threshold for selecting factors, and it can thus consistently estimate the number of above-noise factors. We will state this precisely later.

*4.4.4. Uniqueness of signflips*

The form of Signflip PA suggests a natural generalization: use $H \circ X$ as the "null" data, where $H$ has i.i.d. entries with zero mean and unit variance that are not Rademacher random variables. For example, one might consider using a matrix with i.i.d. standard normal $\mathcal{N}(0,1)$ entries. This raises the question, is there something special about signflips or can anything be used? Might there be a better choice?

From a pragmatic perspective, it is perhaps enough to know that signflips are effective. Especially so, given that signflips have the added practical benefit of being efficient to generate and easy to use. Nevertheless, the prospect of a better or even optimal choice is alluring and is moreover an interesting theoretical question. However, it turns out that signflips are in some sense uniquely suited for deriving theoretical guarantees, as we now describe.

A key step in proving noise recovery for heterogeneous noise in theorem 4.4.11 was proving convergence of the operator norm of $R \circ N$ to the upper-edge of its limiting spectral distribution. We accomplished this by establishing that each of the entries of $R \circ E$ has a sharp sub-Gaussian Laplace transform. This condition is important, and one can refer to the recent works Guionnet and Husson (2020); Husson (2020) for more details. It is not hard to see that $H \circ E$ does not satisfy this assumption in general. For example, suppose both were Gaussian, i.e., $H_{ij} \sim \mathcal{N}(0,1)$ and $E_{ij} \sim \mathcal{N}(0,1)$. Then $H_{ij}E_{ij}$ is the product of two independent standard Gaussians and is no longer sub-Gaussian, let alone sharp sub-Gaussian. In fact, the following proposition shows that Rademacher random variables are the *only* choice for $H_{ij}$ for which $H_{ij}E_{ij}$ has sharp sub-Gaussian Laplace transform when $E_{ij}$ is Gaussian.




(a) Signflipped noise $R \circ N$

(b) Permuted noise $N_\pi$

Figure 23: The empirical and limiting spectral distributions of $R \circ N$ and $N_\pi$, where the first $n/2$ samples have entries with variance $1/(10n)$ and the remainder have entries with variance $9/(10n)$. Limiting spectral distributions shown on top using SPECTRODE (Dobriban, 2015).

**Proposition 4.4.12** (Sharp sub-Gaussianity implies signflips for Gaussian noise)**.** *Let $X \sim \mathcal{N}(0,1)$ be a standard normal and $Y$ be mean zero with unit variance which is independent of $X$. If $XY$ has a sharp sub-Gaussian Laplace transform, then $Y$ must be a Rademacher random variable.*

The proof is given in section 4.8.2. Thus, signflips are uniquely suited for establishing convergence under general noise distributions, at least based on our current theoretical tools. This does not imply that other distributions will necessarily perform poorly, and the opportunity to find better choices remains, e.g., one might try to tailor the choice given certain noise properties. Simply put, other choices fall outside the bounds of our current analysis techniques and would require new approaches to derive guarantees.

*4.4.5. Noise homogenization by permutation*

This section explains why Permutation PA degrades for heterogeneous noise. Consider noise $N = n^{-1/2}(T \circ E) \in \mathbb{R}^{n \times p}$ as in theorem 4.4.11, i.e., $N_{ij}$ are independent with variance $T_{ij}^2/n$. Let $\pi = (\pi_1, \pi_2, \cdots, \pi_p)$ denote the array of independent random permutations used by Permutation PA ($\pi_i$ permutes the entries of the $i$th column). Then one can verify that the marginal variance of $(N_\pi)_{ij}$ is $\tilde{T}_{ij}^2 := (1/n)\sum_{i=1}^n T_{ij}^2$, where the variance is taken with

respect to both $N$ and $\pi$. Namely, $\tilde{T}$ is a homogenized version of $T$ obtained by averaging variances within each column. The permuted noise has a homogenized (marginal) variance profile $\tilde{T}$, so we might expect the spectrum of permuted noise to behave more like a noise matrix with profile $\tilde{T}$ than the actual profile $T$.

Indeed, this intuition plays out in practice, which we illustrate with a simple example in fig. 23. We generate an $n \times p$ noise matrix $N$ with independent normal entries, where $n = 500$ and $p = 300$. The first $n/2$ samples have entries with variance $1/(10n)$, while the remainder have entries with variance $9/(10n)$. This is a piecewise constant variance profile, and indeed signflipped noise accurately recovers the noise spectrum (fig. 23a). On the other hand, the empirical spectral distribution of permuted noise is quite different from that of the noise; permutation significantly shrank the spectrum. This is the general reason Permutation PA suffers under heterogeneous noise. Permutations homogenize the noise, leading to unreliable estimates of the noise level, and consequently an inaccurate selection of the number of factors.

Using SPECTRODE (Dobriban, 2015) we compute and overlay the limiting spectral distributions for random matrices with independent entries and variance profiles $T$ and $\tilde{T}$. Note that $p/n \to \gamma = 3/5$ here, and $\tilde{T}_{ij} = 1/2$ for all $i, j$. The spectrum of $N_\pi$ closely matches the limiting spectral distribution for the profile $\tilde{T}$, even though $N_\pi$ does not actually have independent entries (due to the permutation). This naturally leads one to conjecture that the limiting spectral distribution of $N_\pi$ is nevertheless the same as that for a matrix with independent entries and variance profile $\tilde{T}$. This conjecture is in fact true, as we state in the following theorem that concludes the theoretical explanation by providing a rigorous characterization of the limiting spectrum of permuted noise $N_\pi$ under heterogeneous variance profiles.

**Theorem 4.4.13** (Permutations homogenize variance profiles)**.** *Let $N = n^{-1/2}(T \circ E) \in \mathbb{R}^{n \times p}$, where $T = T_{n,p}$ is a sequence of deterministic variance profiles and $E = E_{n,p} \in \mathbb{R}^{n \times p}$ has independent entries with zero mean, unit variance, and uniformly bounded fourth*

*moments. Suppose $T$ has nonnegative uniformly bounded entries and its column mean squares,*

$$\eta_j^2 := \frac{1}{n}\sum_{i=1}^{n} T_{ij}^2, \quad \textit{for } j = 1, \ldots, p,$$

*have empirical distribution converging to a deterministic distribution $H$. Then as $n, p \to \infty$ with $p/n \to \gamma > 0$, with probability one, the empirical spectral distribution of $N_\pi^\top N_\pi$ for permuted noise $N_\pi$ converges weakly to the generalized Marchenko-Pastur distribution, whose Stieltjes transform $m(z)$ satisfies:*

$$1 + \frac{1}{\gamma(zm(z)+1) - 1} = \int \frac{\gamma t}{\gamma t(zm(z)+1) + z - t} dH(t), \quad z \in \mathbb{C}^+. \tag{4.2}$$

A key special case is when $\eta_1^2 = \cdots = \eta_p^2 = 1$, i.e., all column mean squares of $T$ are unity. In this case, $N_\pi$ has a spectrum like a random matrix with i.i.d. entries.

To prove theorem 4.4.13, we actually first prove the following lemma. It establishes a generalized Marchenko-Pastur law under relaxed independence conditions, and is of independent interest. We allow dependence among entries while imposing conditions on the population covariances of the row. For some other related (but different) results, see Hui and Pan (2010); Wei et al. (2016); Bryson et al. (2019).

**Lemma 4.4.14** (Generalized Marchenko-Pastur with relaxed independence conditions)**.** *Let $X = X_{n,p} \in \mathbb{R}^{n \times p}$ be a sequence of zero mean random matrices with independent rows. Suppose that $n, p \to \infty$ with $p/n \to \gamma > 0$ and furthermore:*

1. *Each row $x_k$ of $X$ has scalar covariance $\mathbb{E}(x_k x_k^\top) = \eta_k^2 I_p$.*
2. *The variances $\eta_1^2, \ldots, \eta_n^2$ are uniformly bounded with empirical distribution converging to some deterministic limiting distribution $H$.*
3. *For any deterministic $p \times p$ matrices $A = A_p$ with uniformly bounded spectral norm and*

*for every row $x_k$, we have*

$$\mathrm{Var}\left(x_k^\top A x_k\right) = o(p^2). \tag{4.3}$$

*Then, with probability one, the empirical spectral distribution of $n^{-1}X^\top X$ converges weakly to the generalized Marchenko-Pastur distribution, whose Stieltjes transform $m(z)$ satisfies:*

$$z + \frac{1}{m(z)} = \int \frac{t}{1+\gamma t m(z)} dH(t), \quad z \in \mathbb{C}^+. \tag{4.4}$$

This lemma is proved in section 4.8.3 by carefully combining techniques used in the proofs of (Bai and Zhou, 2008, Theorem 1.1) and (Bai and Silverstein, 2010, Theorem 4.3). With this lemma in hand, we prove theorem 4.4.13 in section 4.8.4. The key is to show the permuted noise matrix satisfies all the conditions of theorem 4.4.14, of which especially crucial is the concentration of quadratic forms eq. (4.3).

## 4.5. Implications for rank selection

This section explains the implications of the general theoretical characterization for rank selection. In particular, we will prove that Signflip PA is asymptotically accurate, in a sense which we now precisely define.

**Definition 4.5.1** (Asymptotically perceptible factors and accurate selection)**.** *Suppose $X = X_{n,p} = S+N \in \mathbb{R}^{n\times p}$ is a sequence of data matrices for which the noise upper-edge converges in probability: $\|N\| \to \tilde{\sigma}$. Then, as $n, p \to \infty$, we say the $k$-th factor is*

***perceptible (in probability)*** *if $\mathbb{P}\left[\sigma_k(X) > \tilde{\sigma} + \varepsilon\right] \to 1$ for some $\varepsilon > 0$.*

***imperceptible (in probability)*** *if $\mathbb{P}\left[\sigma_k(X) < \tilde{\sigma} - \varepsilon\right] \to 1$ for some $\varepsilon > 0$.*

*In particular, if $\sigma_k(X) \to \hat{\sigma}$ in probability, then it is perceptible if $\hat{\sigma} > \tilde{\sigma}$ and imperceptible if $\hat{\sigma} < \tilde{\sigma}$. We say a selection is* perceptive *if it includes all perceptible factors and no imperceptible factors.*

This notion was introduced in Dobriban (2020) and naturally captures which factors are

asymptotically meaningful and "rise above the noise". Factors falling "below the phase transition" asymptotically tend to produce essentially noise-like factors (see, e.g., Baik et al., 2005; Baik and Silverstein, 2006; Benaych-Georges and Nadakuditi, 2012; Dobriban and Owen, 2018; Dobriban, 2020). While the signal rank may grow in $n, p$, the number of perceptible factors is an asymptotic property of the sequence. Convergence of the noise upper-edge to a noise level $\|N\| \to \tilde{\sigma}$, ensures that our setting has a meaningful asymptotic notion of the noise "level" or "floor". In this setting, we can hope to consistently estimate factors rising above the noise.

We now to bridge this notion with the theoretical characterization of signflips. Related (but slightly different) lemmas appear in Dobriban and Owen (2018); Dobriban (2020).

**Lemma 4.5.2** (Consistency)**.** *Suppose that data $X = X_{n,p} = S + N \in \mathbb{R}^{n\times p}$ has both*

- *Signal destruction: $\|R \circ S\| \to 0$ in probability, and*
- *Sign-invariant noise: for any fixed $k$, $|\sigma_k(R \circ N) - \sigma_k(N)| \to 0$ in probability,*

*where $R = R_{n,p} \in \{\pm 1\}^{n\times p}$ is the corresponding Rademacher random matrix. Then signflips consistently recover each noise singular value, i.e., for any fixed $k$, $|\sigma_k(R\circ X) - \sigma_k(N)| \to 0$ in probability.*

*Additionally, Signflip PA with upper-edge comparison is perceptive (i.e., selects all perceptible factors and no imperceptible ones) with probability tending to one if the noise upper-edge converges, i.e., $\|N\| \to \tilde{\sigma}$ in probability.*

*Signflip PA with pairwise comparison is also perceptive with probability tending to one if each leading noise singular value similarly converges, i.e., for any fixed $k$, $\sigma_k(N) \to \tilde{\sigma}$ in probability.*

The proof is given in section 4.9.1.

*Remark* 7. Both perceptibility (theorem 4.5.1) and consistency (theorem 4.5.2) are *in*

*probability* for convenience in this section, but one can verify that the analogous notion and statement also apply for almost sure convergence.

*Remark* 8. One may wonder why signal destruction is crucial. Roughly speaking, the main adverse effect is a general inflation of the noise singular values which can lead to under-selection as perceptible factors fall below the inflated threshold. This is called *shadowing*, see e.g., Peres-Neto et al. (2005); Dobriban (2020) and references therein.

We illustrate the power of our theoretical characterization for the important setting of factor analysis. In particular, consider the standard linear common-factor model (e.g., Anderson, 2003; Brown, 2014; Dobriban, 2020). Under this model, the $n \times p$ data matrix $X$ can be written as

$$X = H\Lambda^\top + \mathcal{E},$$

where $\Lambda$ is the fixed $p \times r$ factor loading matrix, $H$ is an $n \times r$ random matrix containing the factor values, and $\mathcal{E}$ is the $n \times p$ idiosyncratic noise matrix. In the standard setting, $X$ has i.i.d. rows $x_i = \Lambda\eta_i + \varepsilon_i$ with covariance matrix

$$\Sigma = \Lambda\Psi\Lambda^\top + \Phi.$$

Here $\Phi$ is the diagonal matrix of idiosyncratic variances and $\Psi = \mathrm{Cov}(\eta_i)$ is the covariance matrix of the factor values.

Our main result is that Signflip PA can correctly select the perceptible factors in a *more general factor model*, where the noise $\mathcal{E}$ can have a *general variance profile*. As discussed above, permutation PA will fail in this setting. Before we state our result, it is convenient to define the scaled factor loading matrix $\Lambda\Psi^{1/2} = (f_1, \ldots, f_r)$.

**Theorem 4.5.3** (Main result: Signflip PA selects the perceptible factors in general heterogeneous noise)**.** *Suppose we have a factor model $X = H\Lambda^\top + \mathcal{E}$, where we observe $n$ samples and each sample is a $p$-dimensional vector. Assume the following conditions:*

1. ***Asymptotics.*** *The aspect ratio $p/n \to \gamma > 0$ as $n, p \to \infty$.*

2. ***Factors.*** *The number $r$ of factors can grow with $n, p$. The rows of the $n \times r$ matrix $H$ are the $n$ samples of $r$ factors. The $i$-th row $\eta_i$ has the form $\eta_i = \Psi^{1/2} U_i$, where $U_i$ has $r$ independent sub-Gaussian entries with zero mean and unit variance.*

3. ***Idiosyncratic noises.*** *The noise $\mathcal{E} = T \circ E$ has a general variance profile, where $T$ and $E$ are as in theorem 4.4.11.*

4. ***Factor loadings.*** *The $r$ scaled factor loading vectors $f_k$ are delocalized, in the sense that $\sum_{k=1}^{r} \|f_k\|_\infty \to 0$ and $n^{-1/2}(\log n)^{1/2} \sum_{k=1}^{r} \|f_k\|_2 \to 0$.*

*Then, with probability tending to one, Signflip PA selects all perceptible factors and no imperceptible factors.*

The proof is given in section 4.9.2.

*Remark* 9. For Permutation PA with fixed number of factors $r$ and a more restricted noise setting, Dobriban (2020, Theorem 2.1) provided a sufficient condition on factor loadings of $\|f_k\|_4 \to 0$ which implies $\|f_k\|_\infty \to 0$ and $n^{-1/4}\|f_k\|_2 \to 0$, since $\|f_k\|_\infty \leq \|f_k\|_4$ and $\|f_k\|_2 \leq n^{1/4}\|f_k\|_4$. In contrast, the condition on factor loadings in theorem 4.5.3 reduces to $\|f_k\|_\infty \to 0$ and $n^{-1/2}(\log n)^{1/2}\|f_k\|_2 \to 0$ in this case. Hence, our conditions are *stronger* even in the special case considered in that work.

*Remark* 10. Our work allows growing factors at an almost linear rate (so-called strong factors) $\|f_k\|_2^2 \sim n(\log n)^{-(1+\delta)}$ for any $\delta > 0$, which is a typical model used in econometrics (Bai and Ng, 2002, 2008; Onatski, 2010).

*Remark* 11. Recalling that $\mathbb{E}\|u\|_\infty, \mathbb{E}\|v\|_\infty = O(p^{-1/2}\log^{1/2} p)$ for $u$ and $v$ drawn uniformly from the unit sphere, theorem 4.5.2 also immediately gives simple guarantees for rank selection in spiked models with normalized vectors and heterogeneous noise.

## 4.6. Experiments

This section demonstrates the empirical performance of Signflip PA through numerical simulations on homogeneous and heterogeneous noise, realistically generated chlorine data with heterogeneous sensor noise, and real data from single-cell RNA-sequencing.

### *4.6.1. Simulation with homogeneous noise*

We start with a setting where Permutation PA excels: a rank-one signal in homogeneous noise. Specifically,

$$X = \theta uv^\top + E \in \mathbb{R}^{n\times p},$$

where $u$ and $v$ are unit vectors (drawn uniformly from the unit spheres in $\mathbb{R}^n$ and $\mathbb{R}^p$), and the noise has entries $E_{ij} \overset{iid}{\sim} \mathcal{N}(0, 1/n)$. We use the $\alpha = 95\%$ percentile from $T = 10$ parallel trials, and repeat the experiment for 1000 runs.

Figure 24 shows the average rank selected by Permutation PA and Signflip PA with both selection rules, where $n = 500$ and $p = 300$ and we sweep across signal strengths $\theta$. For small $\theta$, the underlying component is imperceptible and none of these approaches select it. As $\theta$ grows, the methods begin to discover this component. Pairwise comparisons over-select initially due to non-asymptotic random fluctuations in the leading noise singular values. Upper-edge comparison is more conservative and mitigates the over-selection, but detects the component on average slightly later.

Notably, permutations and signflips perform similarly in this case. Since $u$ and $v$ are drawn uniformly from the sphere, they are generically delocalized and are destroyed well by *both* permutations and signflips. Importantly, the noise here is also homogeneous and hence preserved by both permutations and signflips. In conclusion, this experiment demonstrates that signflip PA works as well as the more well known and widely used permutation PA, even in the special case of homogeneous noise where permutation PA is applicable.

### *4.6.2. Simulation with heterogeneous noise*

Now suppose 90% of the samples have noise variance $0.4/n$, where the remaining 10% have noise variance $1/n$ as in section 4.6.1. As we have discussed, such settings are quite natural

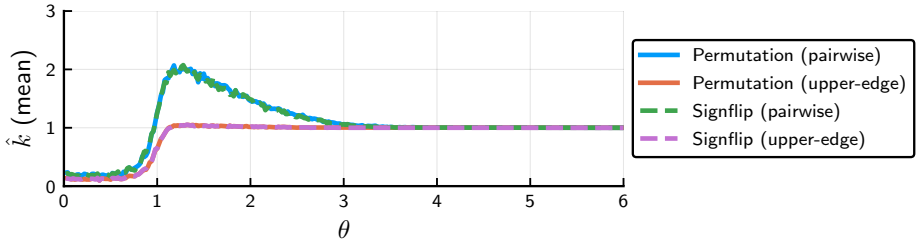



Figure 24: Mean rank selection $\hat{k}$ vs. signal strength $\theta$ for homogeneous noise of variance $1/n$.

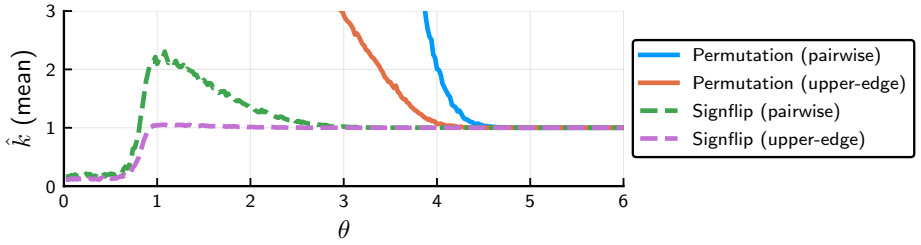
in modern data analysis, e.g., when samples are collected from several instruments or sources that have varying quality. As shown in fig. 25a, Permutation PA now dramatically over-selects. On the other hand, Signflip PA is robust to this heterogeneity and continues to perform well.

In fact, the underlying component is discovered a bit earlier than in fig. 24 (i.e., for smaller $\theta$) since the heterogeneous noise here is actually less noisy overall than the homogeneous noise considered there. Figure 25b compares the scree plots for these two settings. The component is indeed easier to see in the heterogeneous case since the singular value gap is larger.

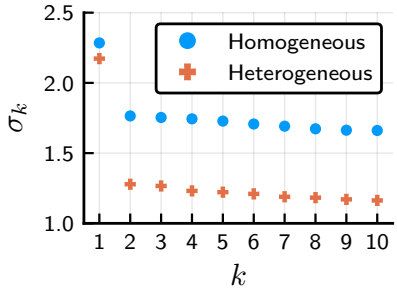
The effect of permutations and signflips on the leading noise singular value, as shown in fig. 25c, corroborate our explanation (section 4.4.5) for why permutations over-select in this presumably easier setting. Permutations homogenize the noise, and as a result the leading singular value of the permuted noise is downwardly biased. Using upper-edge comparison in Permutation PA over-selects, and using pairwise comparison over-selects even more. Signflip PA, on the other hand, recovers the noise (the distribution of the leading noise singular value is essentially the same) and consequently selects the correct number of components.

A natural, practical and (sometimes) effective idea is to normalize the data to make the noise homogeneous.[1] However, this is not always possible. Suppose that 80% of the features have noise variance $0.5/n$ for half the samples and noise variance $1.5/n$ for the other half,

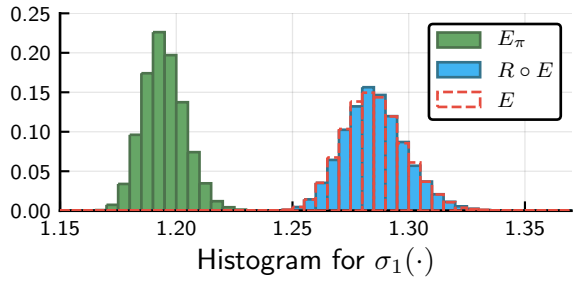
[1]We thank Johannes Heiny and Matthew McKay for raising this point.




(a) Mean rank selection $\hat{k}$ vs. signal strength $\theta$.




(b) Example scree plot for $\theta = 2$.

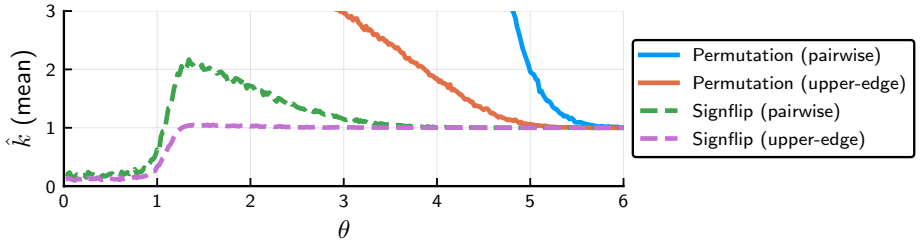



(c) Distribution of leading noise singular value.

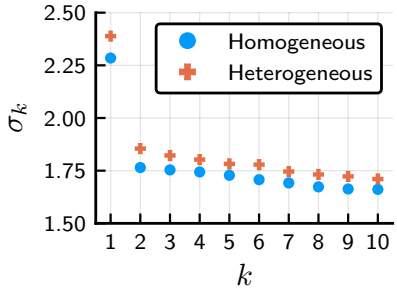
Figure 25: Heterogeneous noise where 90% of samples have noise variance $0.4/n$ and 10% of samples have noise variance $1/n$.

where the other 20% of the features have noise variance $1/n$ for all the samples. This data is normalized in such a way that the features have a *homogeneous* average noise variance of $1/n$. However, the first half of the samples are less noisy than the second, making the sample average noise variances heterogeneous. Unlike fig. 25, re-scaling samples here does not homogenize the noise. Making the sample average noise variances homogeneous would produce heterogeneous feature average noise variances. Such a noise variance profile can arise, e.g., when samples of varying quality on some features are standardized to homogenize feature average noise variances.

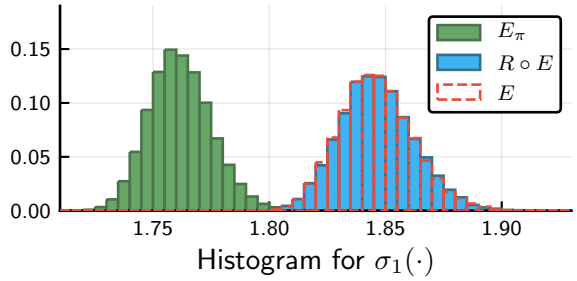
Figure 26 shows the corresponding performance of Permutation PA and Signflip PA for this more complicated heterogeneous example. The overall strength of this noise is more comparable to that of section 4.6.1, so the singular value gaps seen in fig. 26b are also fairly similar. As before, Permutation PA dramatically over-selects (fig. 26a) in this heterogeneous setting since permutations downwardly bias the leading noise singular value (fig. 26c). This underscores that simple normalization methods cannot always address heterogeneity. On




(a) Mean rank selection $\hat{k}$ vs. signal strength $\theta$.




(b) Example scree plot for $\theta = 2$.




(c) Distribution of leading noise singular value.

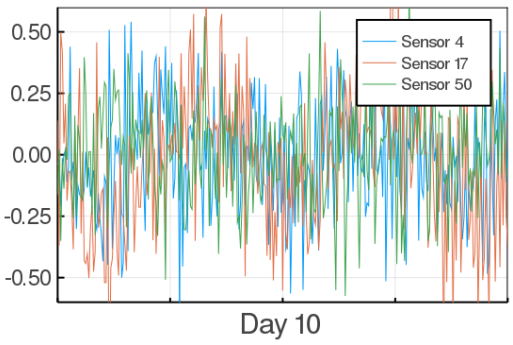
Figure 26: Heterogeneous noise where 80% of features have noise variance $0.5/n$ on half the samples and $1.5/n$ on the second half, and the remaining features have noise variance $1/n$ on all samples.

the other hand, Signflip PA allows for very general noise profiles. In this example, it again accurately recovers the noise and consequently selects the correct number of components. This highlights the flexibility and convenience of using Signflip PA.

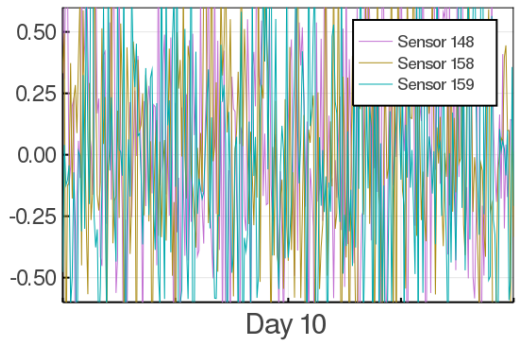
*4.6.3. Realistically generated chlorine data with heterogeneous sensor noise*

This section considers the identification of meaningful components in heterogeneous noise with underlying realistic chlorine measurements generated from EPANET for $n = 166$ sensors at $p = 4310$ time points with a sampling period of five minutes. The data is available online[2] and was previously studied by Papadimitriou et al. (2005); Balzano et al. (2010); as a preprocessing step, we subtract out the DC, i.e., constant, component of each time series. After an initial transient phase, the signals are roughly periodic with a period of approximately 22 hours. To simulate sensors having heterogeneous quality, as can commonly arise in practice, we add mean zero Gaussian noise with variance 0.05 to the first half of the sensors (sensors 1–83) and with variance 0.2 to the second half (sensors 84–166). Figure 27a

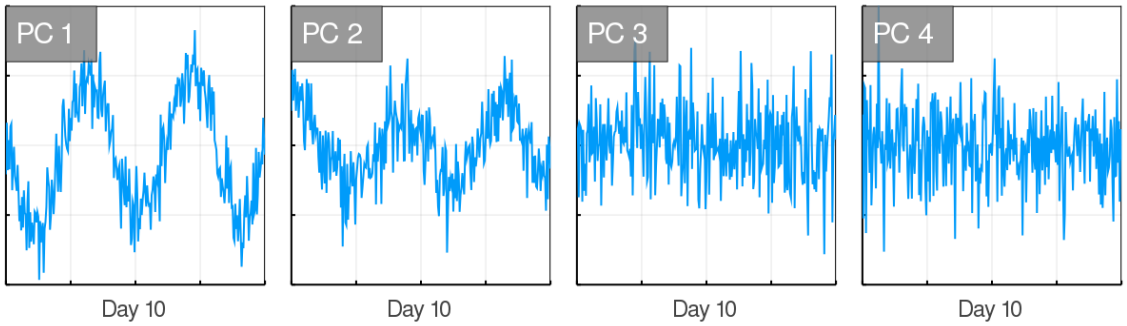
[2]http://www.cs.cmu.edu/afs/cs/project/spirit-1/www/



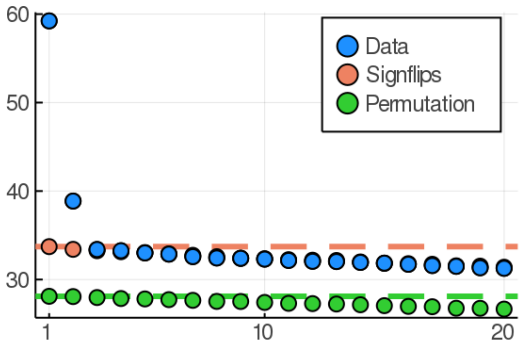

(a) One-day window of a few cleaner sensors.



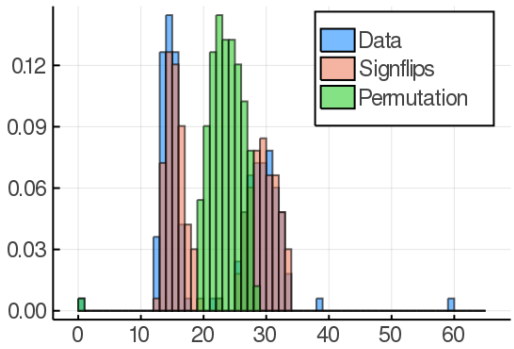

(b) One-day window of a few noisier sensors.




(c) One-day window of first four principal components from noisy curves. The first two components identify the underlying periodic behavior.




(d) Scree plots.




(e) Empirical spectral distributions.

Figure 27: For heterogeneous noise added to realistically generated chlorine data, the components selected by signflip PA appear to rise above the noise and capture meaningful underlying latent behavior. On the other hand, permutations homogenize the noise and select many noise-like components.

shows a few of the cleaner sensors for a one-day window of the data, and fig. 27b shows a few of the noisier sensors for the same window.

Figure 27c shows the leading four right singular vectors of the $n \times p$ mean-centered data matrix formed from the $n$ time series $x_1, \ldots, x_n \in \mathbb{R}^p$ as

$$\bar{X} := \begin{bmatrix} x_1^\top - \bar{x}^\top \\ \vdots \\ x_n^\top - \bar{x}^\top \end{bmatrix}, \quad \text{where } \bar{x} := \frac{1}{n} \sum_{i=1}^{n} x_i.$$

The first two components here appear to rise above the noise and capture the underlying periodic behavior. The remaining components generally seem to have more noise than signal. Roughly speaking, they are below the noise floor and are likely to be essentially random.

The scree plot in fig. 27d shows the first 20 singular values of $\bar{X}$, along with their analogues after permutation and signflipping; a horizontal dashed line indicates the upper-edge comparison cut-off. Signflip PA selects the first two components, consistent with the observation above that these appear to have risen above the noise. Permutation PA, on the other hand, selects many more components in this example. The heterogeneous noise is homogenized by permutations, yielding a different spectrum as shown in fig. 27e. In contrast, signflips preserve the noise spectrum, and consequently appropriately select components that rise above the noise. Notably, signflips here even recover the approximate separation of the bulk singular values into two parts.

*4.6.4. Single-cell RNA-sequencing data*

This section applies Permutation PA and Signflip PA (as well as several other popular non-PA methods) to real data from single-cell RNA-sequencing (scRNA-seq). We use the data set from Macosko et al. (2015). This new sequencing technology has been a powerful tool in the sciences for quantifying the transcriptome of individual cells (Andrews and Hemberg, 2018). The data from scRNA-seq experiments is usually high dimensional and noisy, which

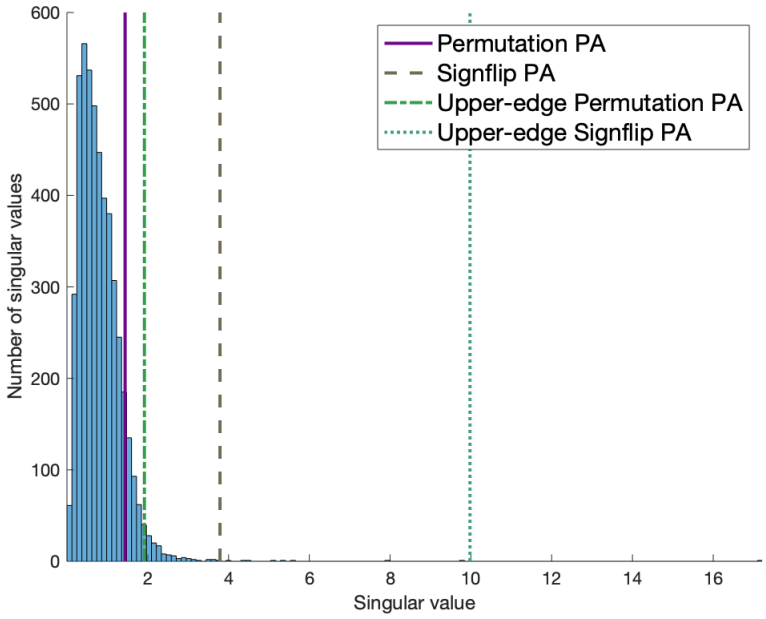



Figure 28: Empirical spectral distribution of the single-cell RNA-sequencing data with the cut-offs chosen by Permutation PA and Signflip PA (with both selection rules).

produces many challenges for data analysis. Moreover, heterogeneous noise may naturally arise due to heterogeneity among cells or genes. Dimensionality reduction, e.g., by PCA or t-SNE, is an important part of data processing pipelines, particularly for data visualization. It is important to determine how many principal components to keep.

The data has $n = 10,000$ cells, and we include $p = 5,000$ genes from each cell. We preprocess the $n \times p$ data matrix by subtracting the mean from each row and normalizing each column. Furthermore, we impute the missing values as zeros. Figure 28 shows the empirical spectral distribution of this data (i.e., the histogram of its singular values) and the cut-offs chosen by Permutation PA and Signflip PA (with both pairwise comparison and upper-edge comparison). We use $T = 20$ trials and a percentile of $\alpha = 100\%$ for all four methods. Permutation PA with pairwise comparison selects 491 components, and using upper-edge comparison selects 128 components. In contrast, Signflip PA with pairwise comparison selects nine, and using upper-edge comparison selects only one. Upper-edge comparison is noticeably more conservative here.

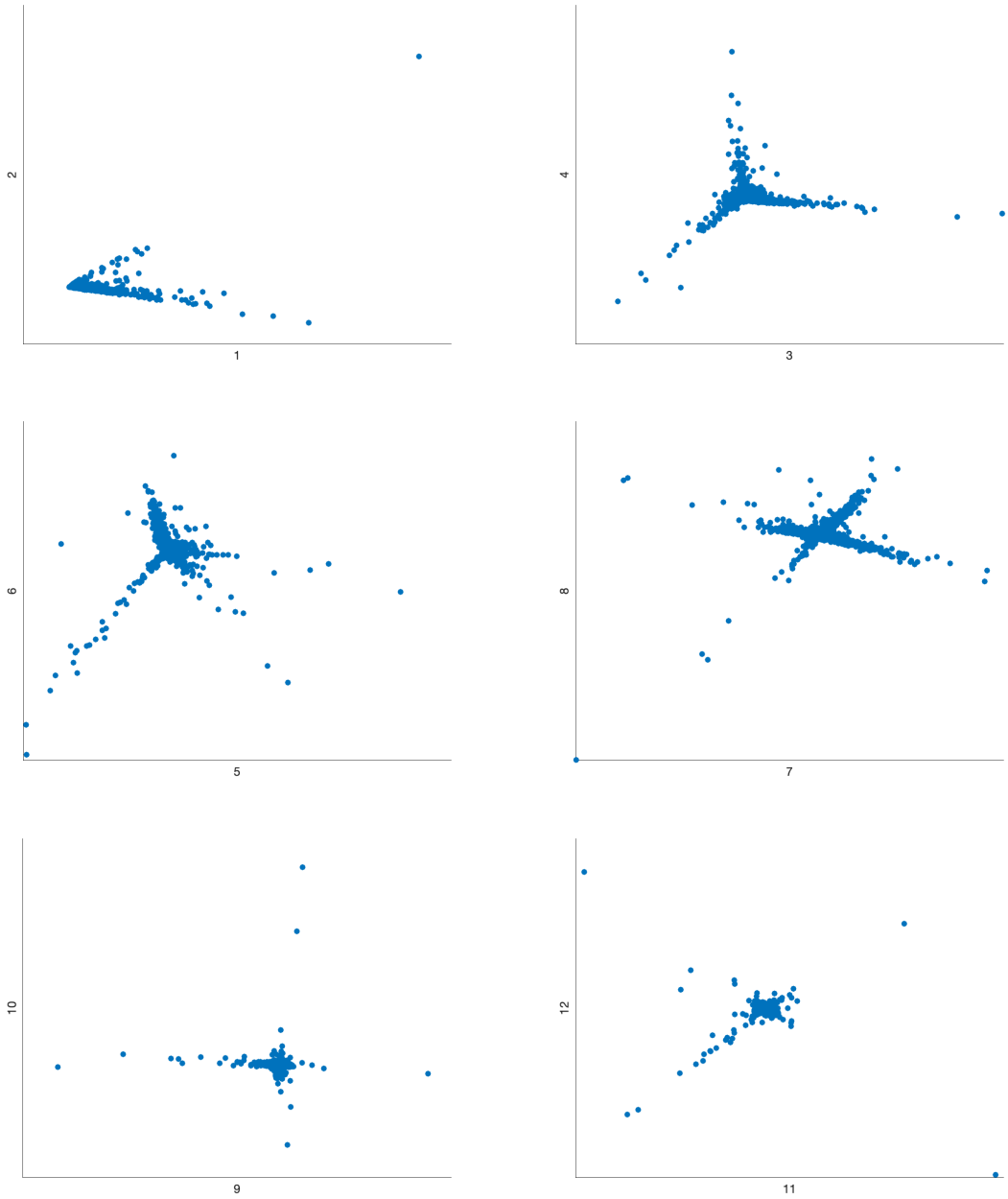



Figure 29: Scatter plots of the first twelve left singular vectors of the scRNA-seq data.

We do not know the ground truth for this real data, but the empirical spectral distribution does appear to have several isolated singular values outside of a "bulk". In particular, the nine singular values above the cut-off shown for Signflip PA with pairwise comparison. We plot the first twelve left singular vectors of the data in fig. 29; each point corresponds to a cell. There seems to be some clustering structure in at least the first eight principal components, which somewhat supports the selection made by Signflip PA with pairwise comparison. The selection of one component by Signflip PA with upper-edge comparison is likely overly conservative, and might be the result of insufficient signal destruction leading to an inflated operator norm estimate. Permutation PA with either selection rule is likely selecting too many principal components (the cut-offs are well into the "bulk"). This may be due to heterogeneity in the noise, highlighting the need for flexible approaches that can accommodate potential heterogeneity in the noise.

We also tried several popular non-PA factor selection methods. The eigenvalue ratio method (Ahn and Horenstein, 2013) selected one factor, while the eigenvalue difference method (Onatski, 2010) selected three components. These methods were designed to tackle models with strong factors, so it is perhaps natural that they tend to select relatively fewer components. The information criterion-based estimator (Nadakuditi and Edelman, 2008) selected 2432 components. This method has theoretical guarantees for identifying weak factors, but under a white noise model. A potential reason why it selects so many components (nearly half) may be due to heterogeneity in the noise.

Indeed, much more work is needed to thoroughly assess the performance of these approaches for a suite of real scRNA-seq datasets, which also have additional complexities (e.g., dependent noise). Nevertheless, in this preliminary experiment, we find that Signflip PA appears to be a promising approach. It gave a reasonable estimate of the rank, with its ability to handle heterogeneous noise potentially aiding the selection. Further investigation is beyond the scope of this paper but is an exciting direction for future work!

## 4.7. Results on signals

In this section, we will prove all remaining results that concern the signals in the signal-plus-noise model. For $L^1$ convergence, we only need to show $\mathbb{E}\|R \circ S\| \to 0$ by definition. For almost sure convergence, by the Borel-Cantelli lemma, it is enough to show that $\mathbb{P}\left[\|R \circ S\| > \delta\right]$ is summable for all $\delta > 0$, i.e. $\sum \mathbb{P}\left[\|R \circ S\| > \delta\right] < \infty$. We will use bounds on either $\mathbb{E}\|R \circ S\|$ or $\mathbb{E}\|R \circ S\|_{S_k}^k$. We can leverage many related works on bounding norms of random matrices with independent entries (Seginer, 2000; Latała, 2005; Schuett and Riemer, 2013; Bandeira and Van Handel, 2016; Van Handel, 2017; Latała et al., 2018). Among all the related papers, Latała et al. (2018) provides some of the tightest bounds, since it fully characterizes $\mathbb{E}\|X\|$ and $\mathbb{E}\|X\|_{S_k}^k$ for a large class of random matrices. We will mainly use results from that paper.

The main result in Latała et al. (2018) is Theorem 1.1: Let $G$ be an $n \times n$ symmetric matrix with i.i.d standard Gaussian variables for the upper triangular part and $S$ be an $n \times n$ symmetric deterministic matrix where all the entries are non-negative. For $X = G \circ S$, we have

$$\left(\mathbb{E}\|X\|_{S_k}^k\right)^{1/k} \asymp \left(\sum_i \left(\sum_j S_{ij}^2\right)^{k/2}\right)^{1/k} + \sqrt{k}\left(\sum_i \max_j S_{ij}^k\right)^{1/k} \tag{4.5}$$

for all $2 \leq k < \infty$, and

$$\mathbb{E}\|X\| \asymp \max_i \sqrt{\sum_j S_{ij}^2} + \max_{i,j} S_{ij}^* \sqrt{\log i}. \tag{4.6}$$

Here the matrix $(S_{ij}^*)$ is obtained by permuting the rows and columns of $S$ such that $\max_j S_{1j}^* \geq \max_j S_{2j}^* \geq \cdots \geq \max_j S_{nj}^*$.

Latała et al. (2018), (on page 1036), gives the following probabilistic interpretation (dating back to prior work by Seginer (2000)). The expected operator norm is of the same order as

*the expected largest row norm*:

$$\mathbb{E}\|X\| \asymp \mathbb{E}\max_i \sqrt{\sum_j X_{ij}^2}. \tag{4.7}$$

By symmetry, this is the same as the largest column norm. This result is quite surprising, because it shows that the interaction between the rows/columns is quite weak. Latała et al. (2018) establish that the right hand side has the rate in (4.6).

The theorem above holds for symmetric matrices and we also need the Gaussianity assumption. But we can apply two tricks (e.g., from Latała (2005)) to make both (4.5) and (4.6) useful for our purpose. First, since both $\|\cdot\|$ and $\|\cdot\|_{S_k}^k$ are convex functions on matrices, by using Jensen's inequality, we have the following:

$$\mathbb{E}\|G \circ S\| = \mathbb{E}\||G| \circ R \circ S\| \geq \mathbb{E}_R\|\mathbb{E}_G|G| \circ R \circ S\| = \sqrt{\frac{2}{\pi}}\mathbb{E}\|R \circ S\|$$

and

$$\mathbb{E}\|G \circ S\|_{S_k}^k = \mathbb{E}\||G| \circ R \circ S\|_{S_k}^k \geq \mathbb{E}_R\|\mathbb{E}_G|G| \circ R \circ S\|_{S_k}^k = \left(\frac{2}{\pi}\right)^{k/2} \cdot \mathbb{E}\|R \circ S\|_{S_k}^k.$$

Second, we can apply the block matrix $\tilde{X} = [0, X; X^\top, 0]$ to (4.5) and (4.6). Notice that $\|\tilde{X}\| = \|X\|$ and $\|\tilde{X}\|_{S_k}^k = 2\|X\|_{S_k}^k$. Then (4.5) and (4.6) can be easily extended to rectangular matrices.

We start with the following lemma.

**Lemma 4.7.1** (Expected operator norm of signflipped matrices)**.** *Let $X$ be an $n \times p$ matrix and let $R$ be an $n \times p$ Rademacher random matrix. Then*

$$\begin{aligned}(\mathbb{E}\|R \circ X\|_{S_k}^k)^{1/k} &\lesssim 2^{-1/k}\big\{(\|X\|_{2,k}^k + \|X^\top\|_{2,k}^k)^{1/k} + \sqrt{k}(\|X\|_{\infty,k}^k + \|X^\top\|_{\infty,k}^k)^{1/k}\big\} \\ &\leq 2^{-1/k}\big\{\|X\|_{2,k} + \|X^\top\|_{2,k} + \sqrt{k}(\|X\|_{\infty,k} + \|X^\top\|_{\infty,k})\big\},\end{aligned} \tag{4.8}$$

*for all* $2 \leq k < \infty$, *and*

$$\mathbb{E}\|R \circ X\| \lesssim \max(\|X\|_{2,\infty}, \|X^\top\|_{2,\infty}) + \rho_\infty(X) \leq \|X\|_{2,\infty} + \|X^\top\|_{2,\infty} + \rho_\infty(X), \tag{4.9}$$

*where* $\rho_\infty$ *measures decay of the row/column maximums and is defined as*

$$\rho_\infty(X) := \max_{i=1,\ldots,n+p} \left\| \begin{matrix} & X \\ X^\top & \end{matrix} \right\|_{\infty,(i)} \sqrt{\log i}, \tag{4.10}$$

*where* $\|A\|_{\infty,(i)}$ *denotes the* $i$*-th largest column* $\ell_\infty$ *norm, i.e.,* $\|A\|_{\infty,(1)} \geq \cdots \geq \|A\|_{\infty,(q)}$ *sorts the column norms* $\|A_{:1}\|_\infty, \ldots, \|A_{:q}\|_\infty$ *in descending order for an* $m \times q$ *matrix* $A$.

*4.7.1. Proof of Lemma 4.7.1*

Define the symmetric $(n+p) \times (n+p)$ matrices

$$\tilde{X} := \begin{pmatrix} & X \\ X^\top & \end{pmatrix}, \qquad \tilde{R} := \begin{pmatrix} & R \\ R^\top & \end{pmatrix},$$

and let $\tilde{G}$ be a symmetric $(n+p) \times (n+p)$ Gaussian random matrix, i.e., $\tilde{G}_{ij} \overset{iid}{\sim} \mathcal{N}(0,1)$ for $i \geq j$. Applying (Latała et al., 2018, Theorem 1.1) to $\tilde{G} \circ \tilde{X}$ yields the bounds:

$$\begin{aligned}
(\mathbb{E}\|\tilde{G} \circ \tilde{X}\|_{S_k}^k)^{1/k} &\lesssim \|\tilde{X}\|_{2,k} + \sqrt{k}\|\tilde{X}\|_{\infty,k} && (4.11)\\
&= \big\|\|X^\top\|_{2,k}, \|X\|_{2,k}\big\|_k + \sqrt{k}\big\|\|X^\top\|_{\infty,k}, \|X\|_{\infty,k}\big\|_k \\
&\leq \big\|\|X^\top\|_{2,k}, \|X\|_{2,k}\big\|_1 + \sqrt{k}\big\|\|X^\top\|_{\infty,k}, \|X\|_{\infty,k}\big\|_1, \\
\mathbb{E}\|\tilde{G} \circ \tilde{X}\| &\lesssim \|\tilde{X}\|_{2,\infty} + \max_{i=1,\ldots,n+p} \tilde{X}_{\infty,(i)}\sqrt{\log i} && (4.12)\\
&= \big\|\|X^\top\|_{2,\infty}, \|X\|_{2,\infty}\big\|_\infty + \rho_\infty(X) \\
&\leq \big\|\|X^\top\|_{2,\infty}, \|X\|_{2,\infty}\big\|_1 + \rho_\infty(X),
\end{aligned}$$

where we have used the following identity and bound that holds for all $2 \leq s, t \leq \infty$:

$$\|\tilde{X}\|_{s,t} = \left\| \begin{matrix} & X \\ X^\top & \end{matrix} \right\|_{s,t} = \left\| \|X^\top\|_{s,t}, \|X\|_{s,t} \right\|_t \leq \left\| \|X^\top\|_{s,t}, \|X\|_{s,t} \right\|_1.$$

Next, we bound $R \circ X$ via a comparison that holds for all Schatten-$t$ norms with $2 \leq t \leq \infty$:

$$\begin{aligned} \mathbb{E}\|\tilde{G} \circ \tilde{X}\|_{S_t} &= \mathbb{E}\||\tilde{G}| \circ \tilde{R} \circ \tilde{X}\|_{S_t} = \mathbb{E}_R \mathbb{E}_G \||\tilde{G}| \circ \tilde{R} \circ \tilde{X}\|_{S_t} \\ &\geq \mathbb{E}_R \|\mathbb{E}_G |\tilde{G}| \circ \tilde{R} \circ \tilde{X}\|_{S_t} = \sqrt{\frac{2}{\pi}} \mathbb{E}\|\tilde{R} \circ \tilde{X}\|_{S_t} = 2^{1/t} \sqrt{\frac{2}{\pi}} \mathbb{E}\|R \circ X\|_{S_t}. \end{aligned} \tag{4.13}$$

The first equality holds because $\tilde{G} \circ \tilde{X} =_d |\tilde{G}| \circ \tilde{R} \circ \tilde{X}$, the inequality follows by applying Jensen's inequality to the norm $\|\cdot\|_{S_t}$, and the final two equalities hold since $\mathbb{E}|\tilde{G}_{ij}| = \sqrt{2/\pi}$ and $\|\tilde{R} \circ \tilde{X}\|_{S_t}^t = 2\|R \circ X\|_{S_t}^t$ (recall that the singular values of $\tilde{R} \circ \tilde{X}$ are made of two copies of those of $R \circ X$). This inequality appears, e.g., in Latała (2005). Combining eqs. (4.11) to (4.13) and rewriting yields eqs. (4.8) and (4.9), concluding the proof.

*4.7.2. Proof of Theorem 4.4.1*

Let us first prove the conditions for $L^1$ convergence. For the first condition, we first show the following useful fact: for any matrix $A$, $\|A\| \leq \||A|\|$. It follows from the definition of operator norm:

$$\|A\| = \max_{x \neq 0} \frac{\|Ax\|}{\|x\|} = \frac{\|Au\|}{\|u\|} \leq \frac{\||A||u|\|}{\||u|\|} \leq \max_{x \neq 0} \frac{\||A|x\|}{\|x\|} = \||A|\|.$$

Then, the first condition easily follows from $\mathbb{E}\|R \circ S\| \leq \mathbb{E}\||R \circ S|\| = \mathbb{E}\||S|\|$.

For the second condition, it is enough to show that $\mathbb{E}\|R \circ S\| \lesssim \|S\|_{2,\infty} + \|S^\top\|_{2,\infty} + \rho_\infty(S)$, where we condition on $S$. This readily comes from theorem 4.7.1.

The third condition is a direct consequence of Corollary 4.7 of Bandeira and Van Handel (2016), it also has been proved earlier in Seginer (2000). This condition could be useful in some cases since it comes from a dimension-dependent bound on $\mathbb{E}\|R \circ S\|$, which is different

from the dimension-free bound in Latała et al. (2018).

Next, we will show all the conditions that guarantee $\mathbb{E}\{\rho_\infty(S)\} \to 0$. First, we want to show that, for any $k \geq 2$,

$$\rho_\infty(S) = \max_i \|\tilde{S}\|_{\infty,(i)} \sqrt{\log i} \lesssim \|\tilde{S}\|_{k,k}.$$

A key observation is, borrowing an argument from Latała et al. (2018), that for any $k \geq 2$,

$$i \cdot \|\tilde{S}\|^k_{\infty,(i)} \leq \sum_{m=1}^{i} \|\tilde{S}\|^k_{\infty,(m)} \leq \sum_{m=1}^{n+p} \|\tilde{S}\|^k_{\infty,(m)} = \sum_{m=1}^{n+p} \max_l |\tilde{S}_{ml}|^k \leq \sum_{m=1}^{n+p} \sum_{l=1}^{n+p} |\tilde{S}_{ml}|^k = \|\tilde{S}\|^k_{k,k}.$$

Then, we have

$$\max_i \|\tilde{S}\|_{\infty,(i)} \sqrt{\log i} \leq \max_i \frac{\sqrt{\log i}}{i^{1/k}} \cdot \|\tilde{S}\|_{k,k} \lesssim \|\tilde{S}\|_{k,k},$$

which leads to the following bound

$$\begin{aligned}
\mathbb{E}\|R \circ S\| = \mathbb{E}\|\tilde{R} \circ \tilde{S}\| &\lesssim \max_i \sqrt{\sum_j \tilde{S}^2_{ij}} + \|\tilde{S}\|_{k,k} \\
&= \max\left(\|S\|_{2,\infty}, \|S^\top\|_{2,\infty}\right) + 2^{1/k} \cdot \|S\|_{k,k} \\
&\lesssim \|S\|_{2,\infty} + \|S^\top\|_{2,\infty} + \|S\|_{k,k}.
\end{aligned}$$

Actually, this also follows from prior results, in particular, from the noncommutative Khintchine inequality (Pisier and Xu, 2003). Then, we would like to show $\|S\|_{k,k} \leq \left\{\operatorname{rank}(S)\|S\|_{2,\infty}\|S^\top\|_{2,\infty}\right\}^{1/2}$. For convenience, we will let $k \geq 4$ be an even integer. The following bounds are useful:

$$\|X\|_F^2 \leq \operatorname{rank}(X)\|X\|^2, \quad \|X\|^2 \leq \|X\|_1\|X\|_\infty, \quad \operatorname{rank}(X \circ Y) \leq \operatorname{rank}(X)\operatorname{rank}(Y).$$

We also need two additional bounds:

$$\|X^{\circ k}\|_1 = \max_j \sum_i X^k_{ij} = \sum_i X^k_{ij'} \leq \left(\sum_i X^2_{ij'}\right)^{k/2} \leq \left(\max_j \sum_i X^2_{ij}\right)^{k/2} = \left(\|X\|_{2,\infty}\right)^k.$$

and similarly $\|X^{\circ k}\|_\infty \leq \left(\|X^\top\|_{2,\infty}\right)^k$. Then, the desired result follows from

$$\begin{aligned}\|S\|_{k,k}^k = \|S^{\circ\frac{k}{2}}\|_F^2 &\leq \operatorname{rank}(S^{\circ\frac{k}{2}})\|S^{\circ\frac{k}{2}}\|^2 \leq \operatorname{rank}(S^{\circ\frac{k}{2}})\|S^{\circ\frac{k}{2}}\|_1\|S^{\circ\frac{k}{2}}\|_\infty \\ &\leq \{\operatorname{rank}(S)\}^{k/2}\|S^{\circ\frac{k}{2}}\|_1\|S^{\circ\frac{k}{2}}\|_\infty \leq \left\{\operatorname{rank}(S)\|S\|_{2,\infty}\|S^\top\|_{2,\infty}\right\}^{k/2}.\end{aligned}$$

Finally, by using the Cauchy-Schwarz inequality, we have

$$\mathbb{E}\left\{\operatorname{rank}(S)\|S\|_{2,\infty}\|S^\top\|_{2,\infty}\right\}^{1/2} \leq \sqrt{\operatorname{rank}(S)\mathbb{E}\|S\|_{2,\infty}\mathbb{E}\|S^\top\|_{2,\infty}},$$

which implies $\mathbb{E}\{\rho_\infty(S)\} \to 0$ when $\operatorname{rank}(S)$ is uniformly bounded.

For almost sure convergence, notice that, for any matrix $A$, we have $\|A\|^k \leq \operatorname{tr}(A^\top A)^{k/2}$ for all $k \geq 2$. By Markov's inequality, for any $\delta > 0$, we have

$$\mathbb{P}\left[\|A\| > \delta\right] = \mathbb{P}\left[\|A\|^k > \delta^k\right] \leq \mathbb{P}\left[\operatorname{tr}(A^\top A)^{k/2} > \delta^k\right] \leq \frac{\mathbb{E}\operatorname{tr}(A^\top A)^{k/2}}{\delta^k} = \frac{\mathbb{E}\|R\circ S\|_{S_k}^k}{\delta^k}.$$

So, in order to show $\|R \circ S\| \to 0$ almost surely, we only need to show $\mathbb{E}\|R\circ S\|_{S_k}^k$ is summable for some $k \geq 2$. We condition on $S$ and apply (4.5) to $\tilde{S}$:

$$\begin{aligned}\left(\mathbb{E}\|R\circ S\|_{S_k}^k\right)^{1/k} &= \left(\frac{1}{2}\mathbb{E}\|\tilde{R}\circ\tilde{S}\|_{S_k}^k\right)^{1/k} \lesssim \left(\mathbb{E}\|\tilde{G}\circ\tilde{S}\|_{S_k}^k\right)^{1/k} \\ &\lesssim \left(\sum_i\left(\sum_j \tilde{S}_{ij}^2\right)^{k/2}\right)^{1/k} + \sqrt{k}\left(\sum_i \max_j \tilde{S}_{ij}^k\right)^{1/k} \\ &= \left[\|S\|_{2,k}^k + \|S^\top\|_{2,k}^k\right]^{1/k} + \sqrt{k}\left(\|S^{\circ k}\|_{\infty,1} + \|S^{\circ k,\top}\|_{\infty,1}\right)^{1/k}.\end{aligned}$$

By using inclusion between $l_p$ spaces, i.e. $\|x\|_q \leq \|x\|_p$ holds when $0 < p \leq q$, we have

$$\max_j \tilde{S}_{ij}^k \leq \sum_j \tilde{S}_{ij}^k \leq \left(\sum_j \tilde{S}_{ij}^2\right)^{k/2}.$$

So

$$\|S^{\circ k}\|_{\infty,1} + \|S^{\circ k,\top}\|_{\infty,1} = \sum_i \max_j \tilde{S}_{ij}^k \leq \sum_i \left( \sum_j \tilde{S}_{ij}^2 \right)^{k/2} = \|S\|_{2,k}^k + \|S^\top\|_{2,k}^k,$$

which leads to

$$\mathbb{E}\|R \circ S\|_{S_k}^k \lesssim \mathbb{E}\|S\|_{2,k}^k + \mathbb{E}\|S^\top\|_{2,k}^k.$$

By the Borel-Cantelli lemma, it is sufficient to have the summability of $\mathbb{E}\|S\|_{2,k}^k + \mathbb{E}\|S^\top\|_{2,k}^k$. This finishes the proof.

*4.7.3. Proof of Theorem 4.4.2*

For $L^1$ convergence, plug $S = \theta uv^\top$ in theorem 4.4.1. The sufficient condition we will use is

$$\begin{aligned}\mathbb{E}\|R \circ S\| &\lesssim \mathbb{E}\|S\|_{2,\infty} + \mathbb{E}\|S^\top\|_{2,\infty} + \mathbb{E}\sqrt{\operatorname{rank}(S) \cdot \|S\|_{2,\infty} \cdot \|S^\top\|_{2,\infty}} \\ &\leq \left( \frac{\sqrt{\operatorname{rank}(S)}}{2} + 1 \right) (\mathbb{E}\|S\|_{2,\infty} + \mathbb{E}\|S^\top\|_{2,\infty}).\end{aligned}$$

Since $\operatorname{rank}(S) = 1$, we have $\mathbb{E}\|R \circ S\| \lesssim \mathbb{E}\|S\|_{2,\infty} + \mathbb{E}\|S^\top\|_{2,\infty} = \mathbb{E}\|\theta uv^\top\|_{2,\infty} + \mathbb{E}\|\theta vu^\top\|_{2,\infty} = \theta \cdot \mathbb{E}(\|u\|_\infty + \|v\|_\infty) \to 0$, which is a sufficient condition.

For almost sure convergence, the result follows directly from theorem 4.4.1 by plugging $S = \theta uv^\top$ into the sufficient condition.

*4.7.4. Proof of Corollary 4.4.3*

For $L^1$ convergence, suppose $\mathbb{E}\|u\|_\infty, \mathbb{E}\|v\|_\infty = O(p^{-\alpha_1} \cdot (\log p)^{-\alpha_2})$ and $\theta = O(p^{\beta_1} \cdot (\log p)^{\beta_2})$, from theorem 4.4.2, we have

$$\theta \cdot \mathbb{E}\left(\|u\|_\infty + \|v\|_\infty\right) = O(p^{\beta_1 - \alpha_1} \cdot (\log p)^{\beta_2 - \alpha_2}).$$

To make this quantity go to zero, we need either $\beta_1 < \alpha_1$, or $\beta_1 = \alpha_1$ and $\beta_2 < \alpha_2$.

For almost sure convergence, the following bounds are useful:

$$\|u\|_k^k = \sum_{i=1}^{n} u_i^k \leq \|u\|_\infty^{k-2} \sum_{i=1}^{n} u_i^2 = \|u\|_\infty^{k-2} \quad \text{and} \quad \|v\|_k^k \leq \|v\|_\infty^{k-2}.$$

Now, we have

$$\theta^k \cdot \left( \|u\|_k^k + \|v\|_k^k \right) = O(p^{k\beta_1 - (k-2)\alpha_1} \cdot (\log p)^{k\beta_2 - (k-2)\alpha_2}).$$

Notice that, the series $\sum 1/(p \cdot (\log p)^\kappa)$ converges if $\kappa > 1$, so we need one of the following conditions:

- $\exists k \geq 2$, such that $k\beta_1 - (k-2)\alpha_1 < -1$,
- $\exists k \geq 2$, such that $k\beta_1 - (k-2)\alpha_1 = -1$ and $k\beta_2 - (k-2)\alpha_2 < -1$.

Essentially, the first condition corresponds to the following region in the $\alpha_1 - \beta_1 - \alpha_2 - \beta_2$ plane:

$$\bigcup_{k \in [1, +\infty)} \left\{ (\alpha_1, \beta_1, \alpha_2, \beta_2) \,\middle|\, \beta_1 < \frac{k-2}{k}\alpha_1 - \frac{1}{k} \right\} = \{ (\alpha_1, \beta_1, \alpha_2, \beta_2) \,|\, \beta_1 < \alpha_1 \}.$$

The second condition corresponds to the following region:

$$\bigcup_{k \in [1, +\infty)} \left\{ (\alpha_1, \beta_1, \alpha_2, \beta_2) \,\middle|\, \beta_1 = \frac{k-2}{k}\alpha_1 - \frac{1}{k}, \beta_2 < \frac{k-2}{k}\alpha_2 - \frac{1}{k} \right\}$$

which is equivalent to

$$\left\{ (\alpha_1, \beta_1, \alpha_2, \beta_2) \,\middle|\, \beta_1 < \alpha_1, \frac{2\alpha_2 + 1}{2\alpha_1 + 1} < \frac{\alpha_2 - \beta_2}{\alpha_1 - \beta_1} \right\}.$$

We can see that the second region is a subset of the first one. Thus, the condition for almost sure convergence is $\beta_1 < \alpha_1$.

*4.7.5. Proof of Theorem 4.4.4*

For $L^1$ convergence, we start with the triangular inequality $\mathbb{E}\|R\circ S\| = \mathbb{E}\|R\circ\sum_{i=1}^r \theta_i u_i v_i^\top\| \leq \sum_{i=1}^r \mathbb{E}\|R\circ\theta_i u_i v_i^\top\|$. From the proof of theorem 4.4.2, we have $\mathbb{E}\|R\circ\theta_i u_i v_i^\top\| \lesssim \mathbb{E}\|\theta_i u_i v_i^\top\|_{2,\infty} + \mathbb{E}\|\theta_i v_i u_i^\top\|_{2,\infty} = \theta_i \cdot \mathbb{E}(\|u_i\|_\infty + \|v_i\|_\infty)$. Hence, it is easy to see that $\sum_{i=1}^r \theta_i \cdot \mathbb{E}(\|u_i\|_\infty + \|v_i\|_\infty) \to 0$ is sufficient.

For almost sure convergence, we can first condition on $S$ and use the following bound

$$\begin{aligned}\left(\mathbb{E}\|R\circ S\|_{S_k}^k\right)^{1/k} &\leq \left(\mathbb{E}\left(\sum_{i=1}^r \theta_i \|R\circ(u_i v_i^\top)\|_{S_k}\right)^k\right)^{1/k} \\ &\leq \sum_{i=1}^r \theta_i \left(\mathbb{E}\|R\circ(u_i v_i^\top)\|_{S_k}^k\right)^{1/k} \\ &\lesssim \sum_{i=1}^r \theta_i \left(\|u_i\|_k^k\|v_i\|_2^k + \|u_i\|_2^k\|v_i\|_k^k\right)^{1/k}.\end{aligned}$$

The first inequality comes from the triangular inequality for $\|\cdot\|_{S_k}$, i.e. $\|R\circ S\|_{S_k} \leq \sum \|R\circ(u_i v_i^\top)\|_{S_k}$. The second inequality is a consequence of the Minkowski's inequality: for any two random variables $X$ and $Y$, we have $(\mathbb{E}|X+Y|^p)^{1/p} \leq (\mathbb{E}|X|^p)^{1/p} + (\mathbb{E}|Y|^p)^{1/p}$ for all $p \geq 1$. The last inequality comes from the proof of theorem 4.4.1.

*4.7.6. Proof of Corollary 4.4.5*

For $L^1$ convergence, notice that, $\sum_{i=1}^r \theta_i \mathbb{E}\|u_i\|_\infty = O(p^{\nu_1+\beta_1-\alpha_1}\log^{\nu_2+\beta_2-\alpha_2} p)$ and $\sum_{i=1}^r \theta_i \mathbb{E}\|v_i\|_\infty = O(p^{\nu_1+\beta_1-\alpha_1}\log^{\nu_2+\beta_2-\alpha_2} p)$, so the condition $\sum_{i=1}^r \theta_i \mathbb{E}(\|u_i\|_\infty + \|v_i\|_\infty) \to 0$ (from theorem 4.4.4) corresponds to $\nu_1+\beta_1 < \alpha_1$, or $\nu_1+\beta_1 = \alpha_1$ and $\nu_2+\beta_2 < \alpha_2$.

For almost sure convergence, by using Hölder's inequality, we have

$$\sum_{i=1}^r \theta_i \left(\|u_i\|_k^k + \|v_i\|_k^k\right)^{1/k} \leq \left(\sum_{i=1}^r (\|u_i\|_k^k + \|v_i\|_k^k)\right)^{1/k} \left(\sum_{i=1}^r \theta_i^{k/(k-1)}\right)^{1-1/k}.$$

Then, by using the assumptions, we have

$$\left[\sum_{i=1}^r \theta_i \left(\|u_i\|_k^k + \|v_i\|_k^k\right)^{1/k}\right]^k = O(p^{k(\beta_1+\nu_1)-(k-2)\alpha_1}\log^{k(\beta_2+\nu_2)-(k-2)\alpha_2} p).$$

From the proof of theorem 4.4.3, in order to make it summable, we know that this is equivalent to $\nu_1 + \beta_1 < \alpha_1$.

*4.7.7. Proof of theorem 4.4.6*

The results in this section are standard properties of matrix norms, but we provide them here for completeness and for the reader's convenience. We begin with $\|S\|_{2,\infty}$ and $\|S^\top\|_{2,\infty}$. First, both examples are entrywise matrix norms, so they satisfy the first property. We only need to verify the second property. This follows easily from the definition of $\|\cdot\|_{2,\infty}$:

$$\|X\|_{2,\infty} = \max_j \|X_{:,j}\|_2 \leq \max_{\|u\|_2=\|v\|_2=1} u^\top X v = \|X\|,$$

and similarly $\|X^\top\|_{2,\infty} \leq \|X\|$.

Next, it is sufficient to show that all the quantities are upper bounded by $\|S\|_{2,\infty}$ or $\|S^\top\|_{2,\infty}$.

For $\|S\|_{\infty,\infty}$, we have

$$\|S\|_{\infty,\infty} = \max_{i,j} |S_{ij}| \leq \max_j \left(\sum_{i=1}^n S_{ij}^2\right)^{1/2} = \|S\|_{2,\infty}.$$

For $\frac{1}{\sqrt{n}}\|S\|_F$ and $\frac{1}{\sqrt{p}}\|S\|_F$, we have

$$\frac{1}{n}\|S\|_F^2 = \frac{1}{n}\sum_{i,j} S_{ij}^2 \leq \max_i \left(\sum_{j=1}^p S_{ij}^2\right) = \|S^\top\|_{2,\infty}^2,$$

and similarly $\frac{1}{p}\|S\|_F^2 \leq \|S\|_{2,\infty}^2$. Finally, for $\frac{1}{\sqrt{n}}\|S\|_1$ and $\frac{1}{\sqrt{p}}\|S\|_\infty$, by the Cauchy-Schwarz inequality, we have

$$\|S\|_1 = \max_j \sum_{i=1}^n |S_{ij}| = \sum_{i=1}^n |S_{ij'}| \leq \sqrt{n}\left(\sum_{i=1}^n S_{ij'}^2\right)^{1/2} \leq \sqrt{n}\max_j \left(\sum_{i=1}^n S_{ij}^2\right)^{1/2} = \sqrt{n}\|S\|_{2,\infty},$$

and similarly $\|S\|_\infty \leq \sqrt{p}\|S^\top\|_{2,\infty}$. Finally, the desired result follows by applying theorem 4.4.7.

*4.7.8. Proof of Corollary 4.4.8*

Since $\|S\|_F^2 = \sum_{i=1}^r \theta_i^2$, we have $\sum_{i=1}^r \theta_i^2 / \min(n,p) \to 0$, which follows from theorem 4.4.6.

For the other conditions, when the rank is one, they easily follow from theorem 4.4.7, since we have $\|S\|_{2,\infty} = \|\theta uv^\top\|_{2,\infty} = \theta\|v\|_\infty$ and $\|S^\top\|_{2,\infty} = \|\theta vu^\top\|_{2,\infty} = \theta\|u\|_\infty$. For the rank-$r$ case, we have the following bounds

$$\begin{aligned}\|S\|_{2,\infty}^2 &= \max_j \|S_{:,j}\|_2^2 = \max_j \|\sum_{i=1}^r \theta_i u_i (v_i)_j\|_2^2 = \max_j \sum_{i=1}^r \theta_i^2 (v_i)_j^2 \|u_i\|_2^2 \\ &= \max_j \sum_{i=1}^r \theta_i^2 (v_i)_j^2 \geq \max_j \theta_k^2 (v_k)_j^2 = \theta_k^2 \|v_k\|_\infty^2,\end{aligned}$$

and similarly $\|S^\top\|_{2,\infty}^2 \geq \theta_k^2\|u_k\|_\infty^2$ for all $k$. The desired result follows.

## 4.8. Proofs regarding the noise

### *4.8.1. Proof of Theorem 4.4.11*

First, the global law of $N$ follows from Theorem 3.14 of Couillet and Debbah (2011), since the existence of the l.s.d. of $N$ is equivalent to the existence of the l.s.d. of $N^\top N$. We only need to notice that the existence of the sharp sub-Gaussian Laplace transform already implies the existence of all moments. Since the entries of $R \circ N$ still satisfy all the conditions, we also have the global law of $R \circ N$, and the limiting spectral distributions of $N$ and $R \circ N$ are the same.

For convergence of the leading singular values, we can leverage results from the recent work (Husson, 2020). We can apply Theorem 1.3 of Husson (2020) to $\tilde{N} = [0, N; N^\top, 0]$, since the largest singular value of $N$ is exactly the same as the largest eigenvalue of the generalized Wigner matrix $\tilde{N}$. A tricky point is, the variance profile of $\tilde{N}$ may not satisfy the assumptions directly. However, this issue can be resolved by using approximations (see Lemma 6.4 of Husson (2020) and the discussion thereafter). Finally, it is not hard to see that the leading singular values of $R \circ N$ also converge to the same limit.

### *4.8.2. Proof of Proposition 4.4.12*

The key is to characterize the moment generating function of $Y^2$. By Jensen's inequality, we always have the lower bound $\mathbb{E}e^{t^2Y^2/2} \geq e^{\mathbb{E}(t^2Y^2/2)} = e^{t^2/2}$. On the other hand, since

$XY$ has sharp sub-Gaussian Laplace transform, we have the upper bound

$$e^{t^2/2} = e^{t^2\mathbb{E}(XY)^2/2} \geq \mathbb{E}e^{tXY} = \mathbb{E}_Y\mathbb{E}_X e^{tXY} = \mathbb{E}_Y e^{(tY)^2/2} = \mathbb{E}e^{t^2Y^2/2},$$

where we use the fact that $\mathbb{E}(XY)^2 = 1$, independence of $X$ and $Y$, and the moment generating function for the standard normal $\mathbb{E}e^{tX} = e^{t^2/2}$. Thus, $\mathbb{E}e^{t^2Y^2/2} = e^{t^2/2}$. Taken with $\mathbb{E}Y = 0$ and $\mathbb{E}Y^2 = 1$, it follows that $Y$ is a Rademacher random variable.[3]

*4.8.3. Proof of Lemma 4.4.14*

Following the proof of Theorem 1.1 of Bai and Zhou (2008), we proceed with the proof by the following three steps:

1. $m_n(z) - \mathbb{E}m_n(z) \to 0$, a.s.

2. $\mathbb{E}m_n(z) \to m(z)$, which satisfies eq. (4.4).

3. eq. (4.4) has a unique solution in $\mathbb{C}^+$.

Recall that $m_n(z) = p^{-1}\operatorname{tr}(n^{-1}X^\top X - zI_p)^{-1} = p^{-1}\operatorname{tr}\mathcal{B}_n^{-1}$ where

$$\mathcal{B}_n := \mathbf{B}_n - zI_p \in \mathbb{C}^{p\times p}, \qquad \mathbf{B}_n := \frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top \in \mathbb{C}^{p\times p},$$
$$\mathcal{B}_{k,n} := \mathbf{B}_{k,n} - zI_p \in \mathbb{C}^{p\times p}, \qquad \mathbf{B}_{k,n} := \frac{1}{n}\sum_{i\neq k} x_i x_i^\top \in \mathbb{C}^{p\times p}.$$

Throughout the proof, for fixed $z$, we will write $z = \Re(z) + i\Im(z) = u + iv$, where $u, v$ are the real and imaginary parts of $z$. Since $z \in \mathbb{C}^+$, we always have $v > 0$. Also, for convenience, we assume all $\eta_k^2$ are uniformly bounded by a universal constant $L$.

**Step 1.** $m_n(z) - \mathbb{E}m_n(z) \to_{a.s.} 0$.

[3] See Problem 26.7 of Billingsley (1995).

Using the shorthand notation $\mathbb{E}_k(\cdot) = \mathbb{E}(\cdot|x_{k+1}, \ldots, x_n)$, we have

$$\begin{aligned} m_n(z) - \mathbb{E}m_n(z) &= \mathbb{E}_0 m_n(z) - \mathbb{E}_n m_n(z) \\ &= \sum_{k=1}^{n} (\mathbb{E}_{k-1} m_n(z) - \mathbb{E}_k m_n(z)) \\ &= \frac{1}{p} \sum_{k=1}^{n} (\mathbb{E}_{k-1} - \mathbb{E}_k)(\operatorname{tr} \mathcal{B}_n^{-1} - \operatorname{tr} \mathcal{B}_{k,n}^{-1}) \\ &= \frac{1}{p} \sum_{k=1}^{n} (\mathbb{E}_{k-1} - \mathbb{E}_k)\nu_k \end{aligned}$$

where $\nu_k := \operatorname{tr} \mathcal{B}_n^{-1} - \operatorname{tr} \mathcal{B}_{k,n}^{-1}$. By using Lemma 2.6 of Silverstein and Bai (1995), we have $|\nu_k| \leq v^{-1}$. So $(\mathbb{E}_{k-1} - \mathbb{E}_k)\nu_k$ forms a bounded martingale difference sequence and applying the Burkholder inequality (Bai and Silverstein, 2010, Lemma 2.12) yields

$$\mathbb{E}|m_n(z) - \mathbb{E}m_n(z)|^q \leq K_q p^{-q} \mathbb{E}\left( \sum_{k=1}^{n} |(\mathbb{E}_{k-1} - \mathbb{E}_k)\nu_k|^2 \right)^{q/2} \leq K_q \left(\frac{2}{v}\right)^q p^{-q/2} \left(\frac{p}{n}\right)^{-q/2},$$

which, for $q > 2$, implies $m_n(z) - \mathbb{E}m_n(z) = p^{-1} \sum_{k=1}^{n} (\mathbb{E}_{k-1} - \mathbb{E}_k)\nu_k \to_{a.s.} 0$ by using the Borel-Cantelli lemma.

**Step 2.** $\mathbb{E}m_n(z) \to m(z)$, which satisfies eq. (4.4).

We first define

$$\mathbf{K} = \frac{1}{n} \sum_{k=1}^{n} \frac{\eta_k^2}{1 + n^{-1} \operatorname{tr} \mathcal{B}_{k,n}^{-1} \eta_k^2} I_p = K \cdot I_p, \quad \tilde{\mathbf{K}} = \frac{1}{n} \sum_{k=1}^{n} \frac{\eta_k^2}{1 + n^{-1} \mathbb{E} \operatorname{tr} \mathcal{B}_n^{-1} \eta_k^2} I_p = \tilde{K} \cdot I_p.$$

These two quantities are very important, they both have some useful properties. Since $v > 0$, it is easy to show that $\Im K, \Im \tilde{K} < 0$, i.e., the imaginary parts are negative, this gives us $\|(\mathbf{K} - zI_p)^{-1}\|, \|(\tilde{\mathbf{K}} - zI_p)^{-1}\| \leq 1/v$. Another useful fact is (Couillet and Debbah, 2011, Corollary 3.1), suppose $m(z)$ is the Stieltjes transform of a measure on $\mathbb{R}$, then for $z \in \mathbb{C}^+$

$$\left| \frac{1}{1 + m(z)} \right| \leq \frac{|z|}{\Im(z)}.$$

This can be applied to the summands of $K$ and $\tilde{K}$.

Now, since

$$(\mathbf{B}_n - zI_p) - (\mathbf{K} - zI_p) = \left\{ \frac{1}{n} \sum_{k=1}^{n} x_k x_k^\top \right\} - \mathbf{K},$$

by using the resolvent identity $A^{-1} - B^{-1} = -A^{-1}(A - B)B^{-1}$, we have

$$\begin{aligned}
&(\mathbf{K} - zI_p)^{-1} - (\mathbf{B}_n - zI_p)^{-1} = (\mathbf{K} - zI_p)^{-1}((\mathbf{B}_n - zI_p) - (\mathbf{K} - zI_p))(\mathbf{B}_n - zI_p)^{-1} \\
&= \left\{ \frac{1}{n} \sum_{k=1}^{n} (\mathbf{K} - zI_p)^{-1} x_k x_k^\top (\mathbf{B}_n - zI_p)^{-1} \right\} - (\mathbf{K} - zI_p)^{-1}\mathbf{K}(\mathbf{B}_n - zI_p)^{-1} \\
&= \left\{ \sum_{k=1}^{n} \frac{(\mathbf{K} - zI_p)^{-1}(n^{-1} x_k x_k^\top)\mathcal{B}_{k,n}^{-1}}{1 + n^{-1} x_k^\top \mathcal{B}_{k,n}^{-1} x_k} \right\} - (\mathbf{K} - zI_p)^{-1}\mathbf{K}(\mathbf{B}_n - zI_p)^{-1}
\end{aligned}$$

where the last line uses the equation

$$\begin{aligned}
x_k^\top \mathcal{B}_n^{-1} &= x_k^\top \mathcal{B}_{k,n}^{-1} - \frac{x_k^\top \mathcal{B}_{k,n}^{-1}(n^{-1} x_k x_k^\top)\mathcal{B}_{k,n}^{-1}}{1 + n^{-1} x_k^\top \mathcal{B}_{k,n}^{-1} x_k} = x_k^\top \mathcal{B}_{k,n}^{-1} - \frac{n^{-1} x_k^\top \mathcal{B}_{k,n}^{-1} x_k}{1 + n^{-1} x_k^\top \mathcal{B}_{k,n}^{-1} x_k} x_k^\top \mathcal{B}_{k,n}^{-1} \\
&= \left\{ 1 - \frac{n^{-1} x_k^\top \mathcal{B}_{k,n}^{-1} x_k}{1 + n^{-1} x_k^\top \mathcal{B}_{k,n}^{-1} x_k} \right\} x_k^\top \mathcal{B}_{k,n}^{-1} = \frac{x_k^\top \mathcal{B}_{k,n}^{-1}}{1 + n^{-1} x_k^\top \mathcal{B}_{k,n}^{-1} x_k}.
\end{aligned}$$

Taking the trace and dividing by $p$ yields

$$\begin{aligned}
&\frac{1}{p} \operatorname{tr}(\mathbf{K} - zI_p)^{-1} - \frac{1}{p} \operatorname{tr} \mathcal{B}_n^{-1} && (4.14) \\
&= \frac{1}{p} \left\{ \sum_{k=1}^{n} \frac{n^{-1} x_k^\top \mathcal{B}_{k,n}^{-1} (\mathbf{K} - zI_p)^{-1} x_k}{1 + n^{-1} x_k^\top \mathcal{B}_{k,n}^{-1} x_k} \right\} - \frac{1}{p} \operatorname{tr}(\mathbf{K} - zI_p)^{-1}\mathbf{K}\mathcal{B}_n^{-1} && (4.15) \\
&= \frac{1}{p} \left\{ \sum_{k=1}^{n} \frac{n^{-1} x_k^\top \mathcal{B}_{k,n}^{-1} (\mathbf{K} - zI_p)^{-1} x_k}{1 + n^{-1} x_k^\top \mathcal{B}_{k,n}^{-1} x_k} \right\} - \frac{1}{p} \left\{ \sum_{k=1}^{n} \frac{n^{-1} \eta_k^2 \operatorname{tr}(\mathbf{K} - zI_p)^{-1} \mathcal{B}_n^{-1}}{1 + n^{-1} \operatorname{tr} \mathcal{B}_{k,n}^{-1} \eta_k^2} \right\} && (4.16) \\
&= \frac{1}{p} \sum_{k=1}^{n} \frac{d_k}{1 + n^{-1} x_k^\top \mathcal{B}_{k,n}^{-1} x_k}, && (4.17)
\end{aligned}$$

where

$$d_k = n^{-1}x_k^\top \mathcal{B}_{k,n}^{-1}(\mathbf{K}-zI_p)^{-1}x_k - n^{-1}\eta_k^2 \operatorname{tr}(\mathbf{K}-zI_p)^{-1}\mathcal{B}_n^{-1}\left(\frac{1+n^{-1}x_k^\top \mathcal{B}_{k,n}^{-1}x_k}{1+n^{-1}\operatorname{tr}\mathcal{B}_{k,n}^{-1}\eta_k^2}\right)$$

$$= d_{k1} + d_{k2} + d_{k3},$$

with

$$d_{k1} = n^{-1}\eta_k^2 \operatorname{tr}(\mathbf{K}-zI_p)^{-1}\mathcal{B}_{k,n}^{-1} - n^{-1}\eta_k^2 \operatorname{tr}(\mathbf{K}-zI_p)^{-1}\mathcal{B}_n^{-1}, \tag{4.18}$$

$$d_{k2} = n^{-1}x_k^\top \mathcal{B}_{k,n}^{-1}(\mathbf{K}-zI_p)^{-1}x_k - n^{-1}\eta_k^2 \operatorname{tr}\mathcal{B}_{k,n}^{-1}(\mathbf{K}-zI_p)^{-1}, \tag{4.19}$$

$$d_{k3} = n^{-1}\operatorname{tr}\left\{(\mathbf{K}-zI_p)^{-1}\eta_k^2\mathcal{B}_n^{-1}\left(1-\frac{1+n^{-1}x_k^\top \mathcal{B}_{k,n}^{-1}x_k}{1+n^{-1}\eta_k^2\operatorname{tr}\mathcal{B}_{k,n}^{-1}}\right)\right\}. \tag{4.20}$$

For $d_{k1}$, by using Lemma 2.6 of Silverstein and Bai (1995) and the fact that $\|(\mathbf{K}-zI_p)^{-1}\| \leq 1/v$, we have

$$|d_{k1}| \leq \frac{\eta_k^2\|(\mathbf{K}-zI_p)^{-1}\|}{nv} \leq \frac{L}{nv^2}. \tag{4.21}$$

For $d_{k2}$, we have

$$|d_{k2}| \leq \frac{1}{nv}\left|x_k^\top \mathcal{B}_{k,n}^{-1}x_k - \eta_k^2\operatorname{tr}\mathcal{B}_{k,n}^{-1}\right|. \tag{4.22}$$

So

$$\mathbb{E}|d_{k2}|^2 \leq \frac{1}{n^2v^2}\mathbb{E}\left|x_k^\top \mathcal{B}_{k,n}^{-1}x_k - \eta_k^2\operatorname{tr}\mathcal{B}_{k,n}^{-1}\right|^2 \to 0, \tag{4.23}$$

where we use the condition $\mathbb{E}(x_kx_k^\top) = \eta_k^2 I_p$ and eq. (4.3).

For $d_{k3}$, by using $\|(\mathbf{K}-zI_p)^{-1}\| \leq 1/v$, $|\operatorname{tr}\mathcal{B}_n^{-1}| \leq p/v$, and $n^{-1}\eta_k^2\operatorname{tr}\mathcal{B}_{k,n}^{-1}$ is a Stieltjes

transform, we have

$$|d_{k3}| \leq \frac{p\eta_k^2}{n^2v^2}\left|\frac{x_k^\top \mathcal{B}_{k,n}^{-1}x_k - \eta_k^2 \operatorname{tr}\mathcal{B}_{k,n}^{-1}}{1+n^{-1}\operatorname{tr}\mathcal{B}_{k,n}^{-1}}\right| \leq \frac{p|z|L}{n^2v^3}\left|x_k^\top \mathcal{B}_{k,n}^{-1}x_k - \eta_k^2 \operatorname{tr}\mathcal{B}_{k,n}^{-1}\right|. \tag{4.24}$$

Thus,

$$\mathbb{E}|d_{k3}|^2 \leq \frac{p^2|z|^2L^2}{n^4v^6}\mathbb{E}\left|x_k^\top \mathcal{B}_{k,n}^{-1}x_k - \eta_k^2 \operatorname{tr}\mathcal{B}_{k,n}^{-1}\right|^2 \to 0. \tag{4.25}$$

Combining these together, we have

$$|\mathbb{E}d_k|^2 = |\mathbb{E}d_{k1} + \mathbb{E}d_{k2} + \mathbb{E}d_{k3}|^2 \leq 3(\mathbb{E}|d_{k1}|^2 + \mathbb{E}|d_{k2}|^2 + \mathbb{E}|d_{k3}|^2) \to 0.$$

Hence, by using the fact $n^{-1}x_k^\top \mathcal{B}_{k,n}^{-1}x_k$ is a Stieltjes transform,

$$\left|\frac{1}{p}\mathbb{E}\left(\operatorname{tr}(\mathbf{K}-zI_p)^{-1} - \operatorname{tr}\mathcal{B}_n^{-1}\right)\right| = \left|\frac{1}{p}\sum_{k=1}^n \mathbb{E}\frac{d_k}{1+n^{-1}x_k^\top \mathcal{B}_{k,n}^{-1}x_k}\right| \leq \frac{|z|}{pv}\sum_{k=1}^n |\mathbb{E}d_k| \to 0. \tag{4.26}$$

Next, we will show $\mathbf{K}$ can be replaced by $\tilde{\mathbf{K}}$, i.e.,

$$\left|\frac{1}{p}\mathbb{E}\left(\operatorname{tr}(\tilde{\mathbf{K}}-zI_p)^{-1} - \operatorname{tr}\mathcal{B}_n^{-1}\right)\right| \to 0. \tag{4.27}$$

It is equivalent to show

$$\begin{aligned}
&\left|\frac{1}{p}\mathbb{E}\left(\operatorname{tr}(\tilde{\mathbf{K}}-zI_p)^{-1}-\operatorname{tr}(\mathbf{K}-zI_p)^{-1}\right)\right| = \left|\frac{1}{p}\mathbb{E}\operatorname{tr}\left((\tilde{\mathbf{K}}-zI_p)^{-1}(\tilde{\mathbf{K}}-\mathbf{K})\operatorname{tr}(\mathbf{K}-zI_p)^{-1}\right)\right| \\
&\leq \frac{1}{v^2}\mathbb{E}\|\tilde{\mathbf{K}}-\mathbf{K}\| \\
&= \frac{1}{n^2v^2}\mathbb{E}\left|\sum_{k=1}^{n}\eta_k^4\frac{\mathbb{E}\operatorname{tr}\mathcal{B}_n^{-1}-\operatorname{tr}\mathcal{B}_{k,n}^{-1}}{(1+n^{-1}\eta_k^2\operatorname{tr}\mathcal{B}_{k,n}^{-1})(1+n^{-1}\eta_k^2\mathbb{E}\operatorname{tr}\mathcal{B}_n^{-1})}\right| \\
&\leq \frac{|z|^2L^4}{n^2v^4}\sum_{k=1}^{n}\mathbb{E}\left|\operatorname{tr}\mathcal{B}_{k,n}^{-1}-\mathbb{E}\operatorname{tr}\mathcal{B}_n^{-1}\right| \\
&\leq \frac{|z|^2L^4}{n^2v^4}\sum_{k=1}^{n}\left(\mathbb{E}\left|\operatorname{tr}\mathcal{B}_n^{-1}-\mathbb{E}\operatorname{tr}\mathcal{B}_n^{-1}\right|+\mathbb{E}\left|\operatorname{tr}\mathcal{B}_{k,n}^{-1}-\operatorname{tr}\mathcal{B}_n^{-1}\right|\right) \\
&\to 0.
\end{aligned}$$

In the last line, we use the bound $\left|\operatorname{tr}\mathcal{B}_{k,n}^{-1}-\operatorname{tr}\mathcal{B}_n^{-1}\right| \leq 1/v$, and the $L_1$ convergence of $\operatorname{tr}\mathcal{B}_n^{-1}-\mathbb{E}\operatorname{tr}\mathcal{B}_n^{-1}$ which comes from Step 1 where we have shown the $L_r$ convergence ($r>2$) of $\operatorname{tr}\mathcal{B}_n^{-1}-\mathbb{E}\operatorname{tr}\mathcal{B}_n^{-1}$ by applying the Burkholder inequality.

We finally reached

$$\frac{1}{p}\mathbb{E}\left(\operatorname{tr}(\tilde{\mathbf{K}}-zI_p)^{-1}-\operatorname{tr}\mathcal{B}_n^{-1}\right) = \left(\frac{1}{n}\sum_{k=1}^{n}\frac{\eta_k^2}{1+n^{-1}\eta_k^2\mathbb{E}\operatorname{tr}\mathcal{B}_n^{-1}}-z\right)^{-1}-\frac{1}{p}\mathbb{E}\operatorname{tr}\mathcal{B}_n^{-1}\to 0. \tag{4.28}$$

For each fixed $z\in\mathbb{C}^+$, $\mathbb{E}m_n(z)=p^{-1}\mathbb{E}\operatorname{tr}\mathcal{B}_n^{-1}$ is a bounded sequence. So, for any subsequence $\{n'\}$, there is a subsubsequence $\{n''\}$ such that $\mathbb{E}m_{n''}(z)$ converges to a limit $m(z)$. Then, from eq. (4.28), by using the assumption on $\sigma_k^2$, $m$ should satisfy the following equation

$$\left(\int\frac{t}{1+\gamma tm}dH(t)-z\right)^{-1}=m. \tag{4.29}$$

Now, it is not hard to see that $\Im(m)>0$. Suppose the solution to eq. (4.29) is unique (we will show it in the next step), $\mathbb{E}m_n(z)$ converges to a limit which is the unique solution to eq. (4.29). Combining Step 1, we have $m_n(z)\to_{a.s.} m(z)$ for any fixed $z\in\mathbb{C}^+$. Finally,

applying a standard argument based on Vitali's convergence theorem (e.g., see the proof of Theorem 2.9 of Bai and Silverstein (2010)) yields $m_n(z) \to_{a.s.} m(z)$ for all $z \in \mathbb{C}^+$, which is the unique solution to eq. (4.29).

**Step 3.** Show eq. (4.4) has a unique solution in $\mathbb{C}^+$.

This step follows immediately as a special case of (Bai and Silverstein, 2010, Proof of Theorem 4.3, Step 3). For the benefit of the reader, we reproduce the proof here taking advantage of some simplifications made available by specializing to our setting.

Suppose we have two solutions $m_1, m_2 \in \mathbb{C}^+$ of equation eq. (4.4) for a fixed $z \in \mathbb{C}^+$. Consider the difference

$$m_1 - m_2 = \frac{1}{\int \frac{t}{1+\gamma t m_1} dH(t) - z} - \frac{1}{\int \frac{t}{1+\gamma t m_2} dH(t) - z} \tag{4.30}$$

$$= \frac{\int \frac{\gamma t^2 (m_1 - m_2)}{(1+\gamma t m_1)(1+\gamma t m_2)} dH(t)}{\left(\int \frac{t}{1+\gamma t m_1} dH(t) - z\right)\left(\int \frac{t}{1+\gamma t m_2} dH(t) - z\right)}. \tag{4.31}$$

Since $m_1 \neq m_2$, by using the Cauchy-Schwarz inequality,

$$1 = \frac{\int \frac{\gamma t^2}{(1+\gamma t m_1)(1+\gamma t m_2)} dH(t)}{\left(\int \frac{t}{1+\gamma t m_1} dH(t) - z\right)\left(\int \frac{t}{1+\gamma t m_2} dH(t) - z\right)} \tag{4.32}$$

$$\leq \frac{\int \frac{\gamma t^2}{|1+\gamma t m_1|^2} dH(t)}{\left|\int \frac{t}{1+\gamma t m_1} dH(t) - z\right|^2} \cdot \frac{\int \frac{\gamma t^2}{|1+\gamma t m_2|^2} dH(t)}{\left|\int \frac{t}{1+\gamma t m_2} dH(t) - z\right|^2}. \tag{4.33}$$

On the other hand, consider the imaginary part of $m_1$,

$$\Im m_1 = \frac{\Im z + \Im m_1 \int \frac{\gamma t^2}{|1+\gamma t m_1|^2} dH(t)}{\left|\int \frac{t}{1+\gamma t m_1} dH(t) - z\right|^2} > \Im m_1 \frac{\int \frac{\gamma t^2}{|1+\gamma t m_1|^2} dH(t)}{\left|\int \frac{t}{1+\gamma t m_1} dH(t) - z\right|^2}, \tag{4.34}$$

which gives us

$$\frac{\int \frac{\gamma t^2}{|1+\gamma t m_1|^2} dH(t)}{\left| \int \frac{t}{1+\gamma t m_1} dH(t) - z \right|^2} < 1. \tag{4.35}$$

This is also true for $m_2$, so we have

$$\frac{\int \frac{\gamma t^2}{|1+\gamma t m_1|^2} dH(t)}{\left| \int \frac{t}{1+\gamma t m_1} dH(t) - z \right|^2} \cdot \frac{\int \frac{\gamma t^2}{|1+\gamma t m_2|^2} dH(t)}{\left| \int \frac{t}{1+\gamma t m_2} dH(t) - z \right|^2} < 1, \tag{4.36}$$

which leads to a contradiction. Hence, we must have $m_1 = m_2$.

*4.8.4. Proof of Theorem 4.4.13*

We first apply theorem 4.4.14 to $X = \sqrt{n} N_\pi^\top$ to obtain the limiting distribution of the e.s.d. of $n \cdot p^{-1} N_\pi N_\pi^\top$. Then, by using that $N_\pi^\top N_\pi$ and $N_\pi N_\pi^\top$ have the same non-zero eigenvalues, the desired result follows.

So, it remains to show $X = \sqrt{n} N_\pi^\top$ satisfies all the conditions in theorem 4.4.14. We will first show that the $k$-th row $x_k = (x_{k1}, \ldots, x_{kn})^\top \in \mathbb{R}^n$ of $X$ has population covariance matrix $\mathbb{E}\left(x_k x_k^\top\right) = \eta_k^2 I_n$. Then, we show that $\mathrm{Var}(x_k^\top A x_k) = o(n^2)$ holds for any $k$ and any deterministic $n \times n$ matrices $A = A_n$ with uniformly bounded spectral norm. By assumption, $n^{-1} \sum_i T_{ik}^2 = \eta_k^2$ holds for all $k$.

We have $\mathbb{E}\left(x_k x_k^\top\right) = \mathbb{E}_{\pi,N}\left(x_k x_k^\top\right)$, where in the second term, the expectation is over both the entries of $N$ and the random permutations $\pi = (\pi_1, \ldots, \pi_p)$. By symmetry (due to uniform random permutations), one can verify that

$$\mathbb{E}_{\pi,N}(x_{ki}^2) = \sum_i \mathbb{E}\left(N_{ik}^2\right) = \frac{1}{n} \sum_i T_{ik}^2 = \eta_k^2$$

and

$$\mathbb{E}_{\pi,N}(x_{ki} x_{kj}) = \frac{n}{n(n-1)} \sum_{i \neq j} \mathbb{E}\left(N_{ik} N_{jk}\right) = 0 \quad \text{for } i \neq j,$$

which leads to $\mathbb{E}\left(x_k x_k^\top\right) = \eta_k^2 I_n$.

For $\mathrm{Var}(x_k^\top A x_k)$, we first notice that

$$(x_k^\top A x_k)^\top = x_k^\top A^\top x_k = \frac{1}{2} x_k^\top (A + A^\top) x_k$$

and

$$\frac{1}{2}\|A + A^\top\| \leq \frac{\|A\| + \|A^\top\|}{2} = \|A\|.$$

So, it is sufficient to show $\mathrm{Var}(x_k^\top A x_k) = o(n^2)$ holds for all symmetric matrices $A$ with uniformly bounded spectral norm. We may write $\mathrm{Var}(x_k^\top A x_k) = \mathbb{E}_{\pi,N}\left[(x_k^\top A x_k)^2\right] - \left[\mathbb{E}_{\pi,N}(x_k^\top A x_k)\right]^2$. We start with

$$\begin{aligned}
&\mathbb{E}_{\pi,N}\left[(x_k^\top A x_k)^2\right] = \mathbb{E}_{\pi,N}\left(\sum_{i,j} A_{ij} x_{ki} x_{kj}\right)^2 \\
&= \sum_{i,j,l,m} A_{ij} A_{lm} \mathbb{E}_{\pi,N}(x_{ki} x_{kj} x_{kl} x_{km}) \\
&= \left(\sum_{i=j=l=m} + \sum_{i=j\neq l=m} + \sum_{i=l\neq j=m} + \sum_{i=m\neq j=l}\right) A_{ij} A_{lm} \mathbb{E}_{\pi,N}(x_{ki} x_{kj} x_{kl} x_{km}) \\
&= \sum_i A_{ii}^2 \mathbb{E}_{\pi,N}(x_{ki}^4) + \sum_{i\neq l} A_{ii} A_{ll} \mathbb{E}_{\pi,N}(x_{ki}^2 x_{kl}^2) + 2\sum_{i\neq j} A_{ij}^2 \mathbb{E}_{\pi,N}(x_{ki}^2 x_{kj}^2).
\end{aligned}$$

Notice that for $i \neq l$, we have

$$\mathbb{E}_{\pi,N}(x_{ki}^2 x_{kl}^2) = \frac{\sum_{i\neq j} T_{ik}^2 T_{lk}^2}{n(n-1)} = \frac{(\sum_i T_{ik}^2)^2 - \sum_i T_{ik}^4}{n(n-1)} = \frac{n^2\eta_k^4 - \sum_i T_{ik}^4}{n(n-1)}.$$

Thus,

$$\mathbb{E}_{\pi,N}\left[(x_k^\top A x_k)^2\right] = \sum_i A_{ii}^2 \mathbb{E}_{\pi,N}(x_{ki}^4) + \sum_{i\neq l} A_{ii} A_{ll} \frac{n^2\eta_k^4 - \sum_q T_{qk}^4}{n(n-1)} + 2\sum_{i\neq j} A_{ij}^2 \frac{n^2\eta_k^4 - \sum_q T_{qk}^4}{n(n-1)}.$$

Similarly, by using $\mathbb{E}_{\pi,N}(x_{ki}^2) = n^{-1}\sum_i T_{ik}^2 = \eta_k^2$ we have

$$\begin{aligned}\left[\mathbb{E}_{\pi,N}(x_k^\top A x_k)\right]^2 &= \left(\sum_i A_{ii}\mathbb{E}_{\pi,N}(x_{ki}^2)\right)^2 \\ &= \sum_i A_{ii}^2\left[\mathbb{E}_{\pi,N}(x_{ki}^2)\right]^2 + \sum_{i\neq l} A_{ii}A_{ll}\mathbb{E}_{\pi,N}(x_{ki}^2)\mathbb{E}_{\pi,N}(x_{kl}^2) \\ &= \sum_i \eta_k^4 A_{ii}^2 + \sum_{i\neq l}\eta_k^4 A_{ii}A_{ll}.\end{aligned}$$

By putting these together, we have

$$\begin{aligned}&\mathrm{Var}(x_k^\top A x_k) = \mathbb{E}_{\pi,N}\left[(x_k^\top A x_k)^2\right] - \left[\mathbb{E}_{\pi,N}(x_k^\top A x_k)\right]^2 \\ &= \sum_i A_{ii}^2\left[\mathbb{E}_{\pi,N}(x_{ki}^4) - \eta_k^4\right] + \sum_{i\neq l} A_{ii}A_{ll}\left[\frac{n^2\eta_k^4 - \sum_q T_{qk}^4}{n(n-1)} - \eta_k^4\right] + 2\sum_{i\neq j} A_{ij}^2\frac{n^2\eta_k^4 - \sum_q T_{qk}^4}{n(n-1)}.\end{aligned}$$

For the first term, by using the assumptions that the entries of $E$ have uniformly bounded fourth moments and the entries of $\eta$ are uniformly bounded, we have

$$\sum_i A_{ii}^2\left[\mathbb{E}_{\pi,N}(x_{ki}^4) - \eta_k^4\right] \lesssim \sum_i A_{ii}^2.$$

For the second term, since $|A_{ii}| = |e_i^\top A e_i| \leq \|A\|$, the diagonal entries of the symmetric matrix $A$ are bounded by the spectral norm $\|A\|$, we have

$$\begin{aligned}\left|\sum_{i\neq l} A_{ii}A_{ll}\left[\frac{n^2\eta_k^4 - \sum_q T_{qk}^4}{n(n-1)} - \eta_k^4\right]\right| &\leq n(n-1)\|A\|^2\left|\frac{n^2\eta_k^4 - \sum_q T_{qk}^4}{n(n-1)} - \eta_k^4\right| \\ &= n(n-1)\|A\|^2\left|\frac{n\eta_k^4 - \sum_q T_{qk}^4}{n(n-1)}\right| \\ &\lesssim \|A\|^2 n.\end{aligned}$$

For the last term, we have

$$\left| \sum_{i \neq j} A_{ij}^2 \frac{n^2 \eta_k^4 - \sum_q T_{qk}^4}{n(n-1)} \right| \lesssim \sum_{i \neq j} A_{ij}^2.$$

Thus,

$$\mathrm{Var}(x_k^\top A x_k) \lesssim \sum_i A_{ii}^2 + \|A\|^2 n + \sum_{i \neq j} A_{ij}^2 = \|A\|_F^2 + \|A\|^2 n \leq 2\|A\|^2 n = o(n^2).$$

This finishes the proof.

## 4.9. Proofs regarding rank selection

### *4.9.1. Proof of Lemma 4.5.2*

We begin with

$$\begin{aligned} |\sigma_k(R \circ X) - \sigma_k(N)| &\leq |\sigma_k(R \circ X) - \sigma_k(R \circ N)| + |\sigma_k(R \circ N) - \sigma_k(N)| \\ &\leq \|R \circ S\| + |\sigma_k(R \circ N) - \sigma_k(N)| \to 0, \end{aligned}$$

where the first line uses triangle inequality and the second line uses Weyl interlacing (e.g., see Chapter 1 of Tao (2012)).

Then, consider Signflip PA with upper-edge comparison. Essentially, we want to show $\mathbb{P}\left[\sigma_k(X) > \|R \circ X\|\right] \to 1$, given $\mathbb{P}\left[\sigma_k(X) > \tilde{\sigma} + \varepsilon\right] \to 1$. Since $\|R \circ X\| = \|R \circ X\| - \|N\| + \|N\|$ converges to $\tilde{\sigma}$ in probability, for any $\varepsilon > 0$, we have $\mathbb{P}\left[|\|R \circ X\| - \tilde{\sigma}| \leq \varepsilon\right] \to 1$, which implies $\mathbb{P}\left[\|R \circ X\| \leq \tilde{\sigma} + \varepsilon\right] \to 1$. Then, the desired result follows. For the same reason, we know Signflip PA with upper-edge comparison selects no imperceptible factors. Similarly, we can show results for Signflip PA with pairwise comparison.

### *4.9.2. Proof of Theorem 4.5.3*

We need to verify conditions in theorem 4.5.2. We first normalize $X$ as $n^{-1/2}X = n^{-1/2}H\Lambda^\top + n^{-1/2}\mathcal{E} = n^{-1/2}U\Psi^{1/2}\Lambda^\top + n^{-1/2}\mathcal{E} = S + n^{-1/2}(T \circ E)$. For the noise part, we can use theorem 4.4.11. The remaining condition for consistency from theorem 4.5.2 is to show

$\|R \circ S\| \to 0$ in probability, for which we only need to show $\mathbb{E}\|R \circ S\| \to 0$. To do this, we can write $S$ as $S = n^{-1/2} \sum_{k=1}^{r} u_k f_k^\top$. Applying theorem 4.4.4, we know that the sufficient conditions are $n^{-1/2} \sum_{k=1}^{r} \mathbb{E}\|u_k\|_2 \|f_k\|_\infty \to 0$ and $n^{-1/2} \sum_{k=1}^{r} \mathbb{E}\|u_k\|_\infty \|f_k\|_2 \to 0$. Two key quantities are $\mathbb{E}\|u_k\|_2$ and $\mathbb{E}\|u_k\|_\infty$. In order to obtain tight upper bounds on these terms, we can use Exercise 2.5.10 and Exercise 3.1.4 of Vershynin (2018): $\mathbb{E}\|u_k\|_2 \lesssim n^{1/2}$ and $\mathbb{E}\|u_k\|_\infty \lesssim (\log n)^{1/2}$. This finishes the proof.

# BIBLIOGRAPHY


A. Agarwal, O. Chapelle, M. Dudik, and J. Langford. A reliable effective terascale linear learning system. *Journal of Machine Learning Research*, 15:1111–1133, 2014.

S. C. Ahn and A. R. Horenstein. Eigenvalue ratio test for the number of factors. *Econometrica*, 81(3):1203–1227, 2013.

L. Alessi, M. Barigozzi, and M. Capasso. Improved penalization for determining the number of factors in approximate factor models. *Statistics & Probability Letters*, 80(23):1806–1813, 2010.

A. Ali, J. Z. Kolter, and R. J. Tibshirani. A continuous-time view of early stopping for least squares regression. In *Proceedings of Machine Learning Research*, volume 89, pages 1370–1378, 2019.

G. W. Anderson, A. Guionnet, and O. Zeitouni. *An Introduction to Random Matrices*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2009.

G. W. Anderson, A. Guionnet, and O. Zeitouni. *An Introduction to Random Matrices*. Number 118 in Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2010.

T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley New York, 2003.

T. S. Andrews and M. Hemberg. Identifying cell populations with scrnaseq. *Molecular Aspects of Medicine*, 59:114 – 122, 2018.

D. W. Apley and J. Shi. A factor-analysis method for diagnosing variability in multivariate manufacturing processes. *Technometrics*, 43(1):84–95, 2001.

J. Bai and S. Ng. Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221, 2002.

J. Bai and S. Ng. *Large dimensional factor analysis*. Now Publishers Inc, 2008.

Z. Bai and J. W. Silverstein. *Spectral analysis of large dimensional random matrices*. Springer Series in Statistics. Springer, 2010.

Z. Bai and W. Zhou. Large sample covariance matrices without independence structures in columns. *Statistica Sinica*, 18(2):425–442, 2008.

Z. Bai, D. Jiang, J.-F. Yao, and S. Zheng. Corrections to LRT on large-dimensional covariance matrix by RMT. *The Annals of Statistics*, 37(6B):3822–3840, 2009.

Z. Bai, D. Jiang, J. feng Yao, and S. Zheng. Testing linear hypotheses in high-dimensional regressions. *Statistics*, 47(6):1207–1223, 2013.

Z. Bai, K. P. Choi, Y. Fujikoshi, et al. Consistency of aic and bic in estimating the number of significant components in high-dimensional principal component analysis. *Annals of Statistics*, 46(3):1050–1076, 2018.

J. Baik and J. W. Silverstein. Eigenvalues of large sample covariance matrices of spiked population models. *Journal of Multivariate Analysis*, 97(6):1382–1408, 2006.

J. Baik, G. Ben Arous, and S. Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Annals of Probability*, 33(5):1643–1697, 2005.

L. Balzano, R. Nowak, and B. Recht. Online identification and tracking of subspaces from highly incomplete information. In *Allerton Conference on Communication, Control, and Computing.*, pages 704–711. IEEE, 2010.

A. S. Bandeira and R. Van Handel. Sharp nonasymptotic bounds on the norm of random matrices with independent entries. *The Annals of Probability*, 44(4):2479–2506, 2016.

M. Banerjee and C. Durot. Removing the curse of superefficiency: an effective strategy for distributed computing in isotonic regression. *arXiv preprint arXiv:1806.08542*, 2018.

M. Banerjee, C. Durot, and B. Sen. Divide and conquer in non-standard problems and the super-efficiency phenomenon. *The Annals of Statistics*, 47(2):720–757, 2019a.

M. Banerjee, C. Durot, and B. Sen. Divide and conquer in nonstandard problems and the super-efficiency phenomenon. *The Annals of Statistics*, 47(2):720–757, 2019b.

M. S. Bartlett. A note on the multiplying factors for various $\chi^2$ approximations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 16(2):296–298, 1954.

H. Battey, J. Fan, H. Liu, J. Lu, and Z. Zhu. Distributed testing and estimation under sparse high dimensional models. *The Annals of Statistics*, 46(3):1352–1382, 2018.

R. Bekkerman, M. Bilenko, and J. Langford. *Scaling up machine learning: Parallel and distributed approaches*. Cambridge University Press, 2011.

F. Benaych-Georges and R. R. Nadakuditi. The singular values and vectors of low rank perturbations of large rectangular random matrices. *Journal of Multivariate Analysis*, 111:120–135, 2012.

T. Bertin-Mahieux, D. P. Ellis, B. Whitman, and P. Lamere. The million song dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*, 2011.

D. P. Bertsekas and J. N. Tsitsiklis. *Parallel and distributed computation: numerical methods*, volume 23. Prentice hall Englewood Cliffs, NJ, 1989.

P. Billingsley. *Probability and Measure*. John Wiley & Sons, 1995.

G. E. Blelloch and B. M. Maggs. Parallel algorithms. In *Algorithms and theory of computation handbook*, pages 25–25. Chapman & Hall/CRC, 2010.

C. Bordenave, S. Coste, and R. R. Nadakuditi. Detection thresholds in very sparse matrix completion. *arXiv preprint arXiv:2005.06062*, 2020.

S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

M. Braverman, A. Garg, T. Ma, H. L. Nguyen, and D. P. Woodruff. Communication lower bounds for statistical estimation problems via a distributed data processing inequality. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 1011–1020. ACM, 2016.

T. A. Brown. *Confirmatory factor analysis for applied research*. Guilford Publications, 2014.

J. Bryson, R. Vershynin, and H. Zhao. Marchenko-pastur law with relaxed independence conditions. *arXiv preprint arXiv:1912.12724*, 2019.

A. Buja and N. Eyuboglu. Remarks on parallel analysis. *Multivariate behavioral research*, 27(4):509–540, 1992.

T. Cai, X. Han, and G. Pan. Limiting laws for divergent spiked eigenvalues and largest non-spiked eigenvalue of sample covariance matrices. *The Annals of Statistics*, 48(3): 1255–1280, 2020.

T. T. Cai and H. Wei. Distributed gaussian mean estimation under communication constraints: Optimal rates and communication-efficient algorithms. *arXiv preprint arXiv:2001.08877*, 2020.

R. B. Cattell. The scree test for the number of factors. *Multivariate Behavioral Research*, 1(2):245–276, 1966.

R. B. Cattell and S. Vogelmann. A comprehensive trial of the scree and kg criteria for determining the number of factors. *Multivariate Behavioral Research*, 12(3):289–325, 1977.

X. Chen and M.-g. Xie. A split-and-conquer approach for analysis of extraordinarily large data. *Statistica Sinica*, pages 1655–1684, 2014.

X. Chen, W. Liu, and Y. Zhang. First-order newton-type estimator for distributed estimation and inference. *arXiv preprint arXiv:1811.11368*, 2018a.

X. Chen, W. Liu, and Y. Zhang. Quantile regression under memory constraint. *arXiv preprint arXiv:1810.08264*, 2018b.

X. Chen, W. Liu, and Y. Zhang. First-order newton-type estimator for distributed estimation and inference. *arXiv preprint arxiv:1811.11368*, 2018c.

X. Chen, W. Liu, and Y. Zhang. Quantile regression under memory constraint. *The Annals of Statistics*, 47(6):3244–3273, 2019.

Y. Chen and X. Li. Determining the number of factors in high-dimensional generalised latent factor models. *arXiv preprint arXiv:2010.02326*, 2020.

C.-T. Chu, S. K. Kim, Y.-A. Lin, Y. Yu, G. Bradski, K. Olukotun, and A. Y. Ng. Map-reduce for machine learning on multicore. In *Advances in neural information processing systems*, pages 281–288, 2007.

R. Couillet and M. Debbah. *Random Matrix Methods for Wireless Communications*. Cambridge University Press, 2011.

R. Couillet and W. Hachem. Analysis of the limiting spectral measure of large random matrices of the separable covariance type. *Random Matrices: Theory and Applications*, 3(04):1450016, 2014.

R. Couillet, M. Debbah, and J. W. Silverstein. A deterministic equivalent for the analysis of correlated mimo multiple access channels. *IEEE Transactions on Information Theory*, 57(6):3493–3514, 2011.

C. Davis. All convex invariant functions of hermitian matrices. *Archiv der Mathematik*, 8 (4):276–278, 1957.

J. Dean and S. Ghemawat. Mapreduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113, 2008.

L. Dicker. Ridge regression and asymptotic minimax estimation over spheres of growing dimension. *Bernoulli, to appear*, 2014a.

L. Dicker and M. Erdogdu. Flexible results for quadratic forms with applications to variance components estimation. *The Annals of Statistics*, 45(1):386–414, 2017.

L. H. Dicker. Variance estimation in high-dimensional linear models. *Biometrika*, 101(2): 269–284, 2014b.

L. H. Dicker and M. A. Erdogdu. Maximum likelihood for variance estimation in high-dimensional linear models. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 159–167. PMLR, 2016.

X. Ding and T. Jiang. Spectral distributions of adjacency and Laplacian matrices of random graphs. *The Annals of Applied Probability*, 20(6):2086 – 2117, 2010.

E. Dobriban. Efficient computation of limit spectra of sample covariance matrices. *Random Matrices: Theory and Applications*, 04(04):1550019, 2015.

E. Dobriban. Permutation methods for factor analysis and pca. *The Annals of Statistics*, 48(5):2824–2847, 2020.

E. Dobriban and S. Liu. Asymptotics for sketching in least squares regression. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

E. Dobriban and A. B. Owen. Deterministic parallel analysis: an improved method for selecting factors and principal components. *Journal of the Royal Statistical Society. Series B (Methodological)*, 81(1):163–183, 2018.

E. Dobriban and Y. Sheng. Distributed linear regression by averaging. *arXiv preprint arxiv:1810.00412*, 2018.

E. Dobriban and Y. Sheng. One-shot distributed ridge regression in high dimensions. *arXiv preprint arXiv:1903.09321*, 2019.

E. Dobriban and S. Wager. High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247–279, 2018.

E. Dobriban, W. Leeb, and A. Singer. Optimal prediction in the linearly transformed spiked model. *arXiv preprint arXiv:1709.03393*, 2017.

E. Dobriban, W. Leeb, and A. Singer. Theoretical justification for exponential family pca. *forthcoming*, 2019.

D. L. Donoho and A. Montanari. High dimensional robust M-estimation: Asymptotic variance via approximate message passing. *arXiv preprint arXiv:1310.7320*, 2013.

J. Duan, X. Qiao, and G. Cheng. Distributed nearest neighbor classification. *arXiv preprint arXiv:1812.05005*, 2018.

J. C. Duchi, A. Agarwal, and M. J. Wainwright. Dual averaging for distributed optimization: Convergence analysis and network scaling. *IEEE Transactions on Automatic control*, 57 (3):592–606, 2011.

J. C. Duchi, M. I. Jordan, M. J. Wainwright, and Y. Zhang. Optimality guarantees for distributed statistical estimation. *arXiv preprint arXiv:1405.0782*, 2014.

N. El Karoui and H. Kösters. Geometric sensitivity of random matrix results: consequences for shrinkage estimators of covariance and related statistical methods. *arXiv preprint arXiv:1105.1404*, 2011.

N. El Karoui, D. Bean, P. J. Bickel, C. Lim, and B. Yu. On robust regression with high-dimensional predictors. *Proc. Natl. Acad. Sci. USA*, 110(36):14557–14562, 2013.

L. Erdős, B. Schlein, and H.-T. Yau. Universality of random matrices and local relaxation flow. *Inventiones mathematicae*, 185(1):75–119, 2011.

J. Fan, Q. Sun, W.-X. Zhou, and Z. Zhu. Principal component analysis for big data. *Wiley StatsRef: Statistics Reference Online*, pages 1–13, 2014.

J. Fan, D. Wang, K. Wang, and Z. Zhu. Distributed estimation of principal eigenspaces. *arXiv preprint arXiv:1702.06488*, 2017.

J. Fan, J. Guo, and S. Zheng. Estimating number of factors by adjusted eigenvalues thresholding. *Journal of the American Statistical Association*, 0(ja):1–33, 2020.

E. Fishler, M. Grosmann, and H. Messer. Detection of signals by information theoretic criteria: general asymptotic performance analysis. *IEEE Transactions on Signal Processing*, 50(5):1027–1036, 2002.

V. L. Girko. *Theory of Stochastic Canonical Equations: Volumes I and II*. Springer Netherlands, 2001.

A. Guionnet and J. Husson. Large deviations for the largest eigenvalue of rademacher matrices. *Annals of Probability*, 48(3):1436–1465, 2020.

Z.-C. Guo, L. Shi, and Q. Wu. Learning theory of distributed regression with bias corrected regularization kernel network. *The Journal of Machine Learning Research*, 18(1):4237–4261, 2017.

W. Hachem, P. Loubaton, and J. Najim. The empirical distribution of the eigenvalues of a gram matrix with a given variance profile. *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, 42(6):649–670, 2006.

W. Hachem, P. Loubaton, and J. Najim. Deterministic equivalents for certain functionals of large random matrices. *The Annals of Applied Probability*, 17(3):875–930, 2007.

W. Hachem, P. Loubaton, J. Najim, et al. A clt for information-theoretic statistics of gram random matrices with a given variance profile. *The Annals of Applied Probability*, 18(6): 2071–2130, 2008.

T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.

F. Hiai and D. Petz. *The semicircle law, free random variables and entropy*, volume 77 of *Mathematical Surveys and Monographs*. American Mathematical Soc., 2006.

D. Hong, L. Balzano, and J. A. Fessler. Towards a theoretical analysis of pca for heteroscedastic data. In *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 496–503, Sept 2016.

D. Hong, L. Balzano, and J. A. Fessler. Asymptotic performance of pca for high-dimensional heteroscedastic data. *Journal of multivariate analysis*, 167:435–452, 2018a.

D. Hong, J. A. Fessler, and L. Balzano. Optimally weighted pca for high-dimensional heteroscedastic data. *arXiv preprint arXiv:1810.12862*, 2018b.

D. Hong, Y. Sheng, and E. Dobriban. Selecting the number of components in pca via random signflips. *arXiv preprint arxiv:2012.02985*, 2020.

J. L. Horn. A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2):179–185, 1965.

J. Hui and G. Pan. Limiting spectral distribution for large sample covariance matrices with m-dependent elements. *Communications in Statistics - Theory and Methods*, 39(6): 935–941, 2010.

X. Huo and S. Cao. Aggregated inference. *Wiley Interdisciplinary Reviews: Computational Statistics*, page e1451, 2018.

J. Husson. Large deviations for the largest eigenvalue of matrices with variance profiles. *arXiv preprint arXiv:2002.01010*, 2020.

J. Jiang. Reml estimation: asymptotic behavior and related topics. *The Annals of Statistics*, 24(1):255–286, 1996.

J. Jiang, C. Li, D. Paul, C. Yang, and H. Zhao. On high-dimensional misspecified mixed model analysis in genome-wide association study. *The Annals of Statistics*, 44(5):2127–2160, 2016.

T. Jiang. Low eigenvalues of laplacian matrices of large random graphs. *Probability Theory and Related Fields*, 153:671–690, 2012.

I. M. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *Annals of Statistics*, 29(2):295–327, 2001.

I. M. Johnstone. High dimensional statistical inference and random matrices. In *International Congress of Mathematicians. Vol. I*, pages 307–333. Eur. Math. Soc., Zürich, 2007.

I. M. Johnstone. Multivariate analysis and Jacobi ensembles: Largest eigenvalue, Tracy–Widom limits and rates of convergence. *The Annals of Statistics*, 36(6):2638 – 2716, 2008.

I. M. Johnstone. Approximate null distribution of the largest root in multivariate analysis. *The Annals of Applied Statistics*, 3(4):1616 – 1633, 2009.

I. M. Johnstone and D. Paul. Pca in high dimensions: An orientation. *Proceedings of the IEEE*, 106(8):1277–1292, 2018.

I. Jolliffe. *Principal component analysis*. Wiley Online Library, 2002.

M. I. Jordan, J. D. Lee, and Y. Yang. Communication-efficient distributed statistical inference. *arXiv preprint arXiv:1605.07689*, 2016.

H. F. Kaiser. The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, 20(1):141–151, 1960.

G. Kapetanios. A New Method for Determining the Number of Factors in Factor Models with Large Datasets. Working Papers 525, Queen Mary University of London, School of Economics and Finance, 2004.

G. Kapetanios. A testing procedure for determining the number of factors in approximate factor models with large datasets. *Journal of Business & Economic Statistics*, 28(3): 397–409, 2010.

T. Ke, Y. Ma, and X. Lin. Estimation of the number of spiked eigenvalues in a covariance matrix by bulk eigenvalue matching analysis. *arXiv preprint arXiv:2006.00436*, 2020.

P. Koutris, S. Salihoglu, D. Suciu, et al. Algorithmic aspects of parallel data processing. *Foundations and Trends® in Databases*, 8(4):239–370, 2018.

S. Kritchman and B. Nadler. Non-parametric detection of the number of signals: Hypothesis testing and random matrix theory. *IEEE Transactions on Signal Processing*, 57(10):3930–3941, 2009.

J. Lacotte, S. Liu, E. Dobriban, and M. Pilanci. Optimal iterative sketching methods with the subsampled randomized hadamard transform. In *Advances in Neural Information Processing Systems*, volume 33, pages 9725–9735. Curran Associates, Inc., 2020.

C. Lam and Q. Yao. Factor modeling for high-dimensional time series: Inference for the number of factors. *The Annals of Statistics*, 40(2):694–726, 2012.

R. Latała. Some estimates of norms of random matrices. *Proceedings of the American Mathematical Society*, 133(5):1273–1282, 2005.

R. Latała, R. van Handel, and P. Youssef. The dimension-free structure of nonhomogeneous random matrices. *Inventiones mathematicae*, 214(3):1031–1080, 2018.

D. N. Lawley. Tests of significance for the latent roots of covariance and correlation matrices. *Biometrika*, 43(1-2):128–136, 1956.

J. D. Lee, Q. Liu, Y. Sun, and J. E. Taylor. Communication-efficient sparse regression. *Journal of Machine Learning Research*, 18(5):1–30, 2017.

W. Leeb. Matrix denoising for weighted loss functions and heterogeneous signals. *arXiv preprint arXiv:1902.09474*, 2019.

W. Leeb and E. Romanov. Optimal spectral shrinkage and pca with heteroscedastic noise. *arXiv preprint arXiv:1811.02201*, 2018.

A. S. Lewis. Convex analysis on the hermitian matrices. *SIAM Journal on Optimization*, 6 (1):164–177, 1996.

R. Li, D. K. Lin, and B. Li. Statistical inference in massive data sets. *Applied Stochastic Models in Business and Industry*, 29(5):399–409, 2013.

S.-B. Lin, X. Guo, and D.-X. Zhou. Distributed learning with regularized least squares. *The Journal of Machine Learning Research*, 18(1):3202–3232, 2017.

Z. Lin, C. Yang, Y. Zhu, J. Duchi, Y. Fu, Y. Wang, B. Jiang, M. Zamanighomi, X. Xu, M. Li, N. Sestan, H. Zhao, and W. H. Wong. Simultaneous dimension reduction and adjustment for confounding variation. *Proceedings of the National Academy of Sciences*, 113(51):14662–14667, 2016.

L. T. Liu, E. Dobriban, A. Singer, et al. *e* pca: High dimensional exponential family pca. *The Annals of Applied Statistics*, 12(4):2121–2150, 2018a.

M. Liu, Z. Shang, and G. Cheng. How many machines can we use in parallel computing for kernel ridge regression? *arXiv preprint arXiv:1805.09948*, 2018b.

Q. Liu and A. T. Ihler. Distributed estimation, information loss and exponential families. In *Advances in Neural Information Processing Systems*, pages 1098–1106, 2014.

S. Liu and E. Dobriban. Ridge regression: Structure, cross-validation, and sketching. *arXiv preprint arXiv:1910.02373*, 2019.

M. Lopes, L. Jacob, and M. J. Wainwright. A more powerful two-sample test in high dimensions using random projection. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

N. A. Lynch. *Distributed algorithms*. Elsevier, 1996.

L. W. Mackey, M. I. Jordan, and A. Talwalkar. Divide-and-conquer matrix factorization. In *Advances in neural information processing systems*, pages 1134–1142, 2011.

E. Z. Macosko, A. Basu, R. Satija, J. Nemesh, K. Shekhar, M. Goldman, I. Tirosh, A. R. Bialas, N. Kamitaki, E. M. Martersteck, J. J. Trombetta, D. A. Weitz, J. R. Sanes, A. K. Shalek, A. Regev, and S. A. McCarroll. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, 2015.

V. A. Marchenko and L. A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mat. Sb.*, 114(4):507–536, 1967.

K. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate analysis*. Academic Press, 1979.

R. Mcdonald, M. Mohri, N. Silberman, D. Walker, and G. S. Mann. Efficient large-scale distributed training of conditional maximum entropy models. In *Advances in Neural Information Processing Systems*, pages 1231–1239, 2009.

C. McKennan. Factor analysis in high dimensional biological data with dependent observations. *arXiv preprint arXiv:2009.11134*, 2020.

M. L. Mehta. *Random Matrices*. Academic Press, New York, 3rd edition, 2004.

S. Mei and A. Montanari. The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv preprint arXiv:1908.05355*, 2019.

A. Müller and M. Debbah. Random matrix theory tutorial–introduction to deterministic equivalents. *TRAITEMENT DU SIGNAL*, 33(2-3):223–248, 2016.

R. R. Nadakuditi and A. Edelman. Sample eigenvalue based detection of high-dimensional signals in white noise using relatively few samples. *IEEE Transactions on Signal Processing*, 56(7):2625–2638, 2008.

A. Nedic and A. Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1):48, 2009.

Y. Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical programming*, 120(1):221–259, 2009.

A. Nica and R. Speicher. *Lectures on the combinatorics of free probability*, volume 13. Cambridge University Press, 2006.

A. Onatski. Determining the number of factors from empirical distribution of eigenvalues. *The Review of Economics and Statistics*, 92(4):1004–1016, 2010.

A. B. Owen and J. Wang. Bi-cross-validation for factor analysis. *Statistical Science*, 31(1): 119–139, 2016.

S. Papadimitriou, J. Sun, and C. Faloutsos. Streaming pattern discovery in multiple time-series. In *Proceedings of the 31st International Conference on Very Large Data Bases (VLDB)*, pages 697–708, 2005.

D. Passemier and J. Yao. Estimation of the number of spikes, possibly equal, in the high-dimensional case. *Journal of Multivariate Analysis*, 127:173–183, 2014.

V. H. Patil, M. Q. McPherson, and D. Friesner. The use of exploratory factor analysis in public health: A note on parallel analysis as a factor retention criterion. *American Journal of Health Promotion*, 24(3):178–181, 2010.

D. Paul and A. Aue. Random matrix theory in statistics: A review. *Journal of Statistical Planning and Inference*, 150:1–29, 2014.

D. Paul and J. W. Silverstein. No eigenvalues outside the support of the limiting empirical spectral distribution of a separable covariance matrix. *Journal of Multivariate Analysis*, 100(1):37–57, 2009.

M. J. Peacock, I. B. Collings, and M. L. Honig. Eigenvalue distributions of sums and products of large random matrices via incremental matrix expansions. *IEEE Transactions on Information Theory*, 54(5):2123–2138, 2008.

J. Pennington and P. Worah. Nonlinear random matrix theory for deep learning. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

P. R. Peres-Neto, D. A. Jackson, and K. M. Somers. How many principal components? stopping rules for determining the number of non-trivial axes revisited. *Computational Statistics & Data Analysis*, 49(4):974–997, 2005.

G. Pisier and Q. Xu. Non-commutative lp-spaces. *Handbook of the geometry of Banach spaces*, 2:1459–1517, 2003.

A. Pourshafeie, C. D. Bustamante, and S. Prabhu. Caring without sharing: Meta-analysis 2.0 for massive genome-wide association studies. *bioRxiv*, page 436766, 2018.

T. Rauber and G. Rünger. *Parallel programming: For multicore and cluster systems*. Springer Science & Business Media, 2013.

J. D. Rosenblatt and B. Nadler. On the optimality of averaging in distributed statistical learning. *Information and Inference: A Journal of the IMA*, 5(4):379–404, 2016.

F. Rubio and X. Mestre. Spectral convergence for a general class of random matrices. *Statistics & Probability Letters*, 81(5):592–602, 2011.

C. Schuett and S. Riemer. On the expectation of the norm of random matrices with non-identically distributed entries. *Electronic Journal of Probability*, 18(29), 2013.

S. R. Searle, G. Casella, and C. E. McCulloch. *Variance components*, volume 391. John Wiley & Sons, 2009.

Y. Seginer. The expected norm of random matrices. *Combinatorics, Probability and Computing*, 9(2):149–166, 2000.

V. I. Serdobolskii. On minimum error probability in discriminant analysis. In *Dokl. Akad. Nauk SSSR*, volume 27, pages 720–725, 1983.

V. I. Serdobolskii. *Multiparametric Statistics*. Elsevier, 2007.

O. Shamir, N. Srebro, and T. Zhang. Communication-efficient distributed optimization using an approximate newton-type method. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32, pages 1000–1008, 2014.

Z. Shang and G. Cheng. Computational limits of a distributed algorithm for smoothing spline. *The Journal of Machine Learning Research*, 18(1):3809–3845, 2017.

C. Shi, W. Lu, and R. Song. A massive data framework for m-estimators with cubic-rate. *Journal of the American Statistical Association*, 113(524):1698–1709, 2018.

J. W. Silverstein and Z. Bai. On the empirical distribution of eigenvalues of a class of large dimensional random matrices. *Journal of Multivariate Analysis*, 54(2):175–192, 1995.

V. Smith, S. Forte, C. Ma, M. Takác, M. I. Jordan, and M. Jaggi. Cocoa: A general framework for communication-efficient distributed optimization. *arXiv preprint arXiv:1611.02189*, 2016.

A. Smola. Course notes on scalable machine learning, 2012.

D. W. Stewart. The application and misapplication of factor analysis in marketing research. *Journal of Marketing Research*, 18(1):51–62, 1981.

C. Subbarao, N. V. Subbarao, and S. N. Chandu. Characterization of groundwater contamination using factor analysis. *Environmental Geology*, 28:175–180, 1996.

B. Szabo and H. van Zanten. Adaptive distributed methods under communication constraints. *arXiv preprint arXiv:1804.00864*, 2018.

R. Tandon, Q. Lei, A. G. Dimakis, and N. Karampatziakis. Gradient coding: Avoiding stragglers in distributed learning. In *International Conference on Machine Learning*, pages 3368–3376, 2017.

T. Tao. *Topics in Random Matrix Theory*, volume 132. American Mathematical Soc., 2012.

T. Tao and V. Vu. Random matrices: Universality of local eigenvalue statistics. *Acta Math.*, 206(1):127–204, 2011.

U. S. Tran and A. K. Formann. Performance of parallel analysis in retrieving unidimensionality in the presence of binary data. *Educational and Psychological Measurement*, 69(1):50–61, 2009.

J. Tsitsiklis, D. Bertsekas, and M. Athans. Distributed asynchronous deterministic and stochastic gradient optimization algorithms. *IEEE transactions on automatic control*, 31 (9):803–812, 1986.

A. M. Tulino and S. Verdú. Random matrix theory and wireless communications. *Communications and Information theory*, 1(1):1–182, 2004.

R. Van Handel. Structured random matrices. In *Convexity and Concentration*, pages 107–156. Springer, 2017.

W. F. Velicer. Determining the number of components from the matrix of partial correlations. *Psychometrika*, 41(3):321–327, 1976.

R. Vershynin. *High-Dimensional Probability*. Cambridge University Press, 2018.

D. V. Voiculescu, K. J. Dykema, and A. Nica. *Free random variables*, volume 1 of *CRM Monograph Series*. American Mathematical Soc., 1992.

S. Volgushev, S.-K. Chao, and G. Cheng. Distributed inference for quantile regression processes. *The Annals of Statistics*, 47(3):1634–1662, 2019a.

S. Volgushev, S.-K. Chao, G. Cheng, et al. Distributed inference for quantile regression processes. *The Annals of Statistics*, 47(3):1634–1662, 2019b.

H. Wang. Factor profiled sure independence screening. *Biometrika*, 99(1):15–28, 2012.

J. Wang, M. Kolar, N. Srebro, and T. Zhang. Efficient distributed learning with sparsity. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3636–3645. JMLR. org, 2017.

X. Wang, Z. Yang, X. Chen, and W. Liu. Distributed inference for linear support vector machine. *Journal of Machine Learning Research*, 20(113):1–41, 2019.

M. Wax and T. Kailath. Detection of signals by information theoretic criteria. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 33(2):387–392, 1985.

M. Wei, G. Yang, and L. Yang. The limiting spectral distribution for large sample covariance matrices with unboundedm-dependent entries. *Communications in Statistics - Theory and Methods*, 45(22):6651–6662, 2016.

H. Wickham. *nycflights13: Flights that Departed NYC in 2013*, 2018. URL `https://CRAN.R-project.org/package=nycflights13`. R package version 1.0.0.

G. Xu, Z. Shang, and G. Cheng. Optimal tuning for divide-and-conquer kernel ridge regression with massive data. In *Proceedings of the 35th International Conference on Machine Learning*, pages 5483–5491, 2018.

K. Yano, Y. Morinaka, F. Wang, P. Huang, S. Takehara, T. Hirai, A. Ito, E. Koketsu, M. Kawamura, K. Kotake, S. Yoshida, M. Endo, G. Tamiya, H. Kitano, M. Ueguchi-Tanaka, K. Hirano, and M. Matsuoka. Gwas with principal component analysis identifies a gene comprehensively controlling rice architecture. *Proceedings of the National Academy of Sciences*, 116(42):21262–21267, 2019.

J. Yao, Z. Bai, and S. Zheng. *Large Sample Covariance Matrices and High-Dimensional Data Analysis*. Cambridge University Press, 2015.

Y. Yao, L. Rosasco, and A. Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007.

M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica. Spark: Cluster computing with working sets. *HotCloud*, 10(10-10):95, 2010.

A. Zhang, T. T. Cai, and Y. Wu. Heteroskedastic pca: Algorithm, optimality, and applications. *arXiv preprint arXiv:1810.08316*, 2018.

Y. Zhang, M. J. Wainwright, and J. C. Duchi. Communication-efficient algorithms for statistical optimization. In *Advances in Neural Information Processing Systems*, pages 1502–1510, 2012.

Y. Zhang, J. Duchi, and M. Wainwright. Divide and conquer kernel ridge regression. In *Conference on Learning Theory*, pages 592–617, 2013a.

Y. Zhang, J. C. Duchi, and M. J. Wainwright. Communication-efficient algorithms for statistical optimization. *Journal of Machine Learning Research*, 14:3321–3363, 2013b.

Y. Zhang, J. Duchi, and M. Wainwright. Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. *The Journal of Machine Learning Research*, 16(1):3299–3340, 2015.

T. Zhao, G. Cheng, and H. Liu. A partially linear framework for massive heterogeneous data. *Annals of statistics*, 44(4):1400, 2016.

Y.-H. Zhou. A note on cyclic shift permutation testing for large eigenvalues. *Stat*, 8(1): e257, 2019.

Y. Zhu and J. Lafferty. Distributed nonparametric regression under communication constraints. *arXiv preprint arXiv:1803.01302*, 2018.

M. Zinkevich, J. Langford, and A. J. Smola. Slow learners are fast. In *Advances in neural information processing systems*, pages 2331–2339, 2009.

M. Zinkevich, M. Weimer, L. Li, and A. J. Smola. Parallelized stochastic gradient descent. In *Advances in neural information processing systems*, pages 2595–2603, 2010.